

MEASUREMENT PROPERTIES OF  
RESPONDENT-DEFINED RATING-SCALES

An Investigation of Individual Characteristics and  
Respondent Choices

Elisa CHAMI-CASTALDI

Submitted for the degree of  
Doctor of Philosophy

School of Management

University of Bradford

2010

# MEASUREMENT PROPERTIES OF RESPONDENT-DEFINED RATING-SCALES

## An Investigation of Individual Characteristics and Respondent Choices

Elisa CHAMI-CASTALDI

### Key words:

rating-scales, response categories, response bias, individualised rating-scale procedure (IRSP)

### Abstract:

It is critical for researchers to be confident of the quality of survey data. Problems with data quality often relate to measurement method *design*, through choices made by researchers in their creation of standardised measurement instruments. This is known to affect the way respondents interpret and respond to these instruments, and can result in substantial measurement error. Current methods for removing measurement error are post-hoc and have been shown to be problematic. This research proposes that innovations can be made through the creation of measurement methods that take respondents' individual cognitions into consideration, to reduce measurement error in survey data. Specifically, the aim of the study was to develop and test a measurement instrument capable of having respondents individualise their own rating-scales. A mixed methodology was employed. The qualitative phase provided insights that led to the development of the Individualised Rating-Scale Procedure (IRSP). This electronic measurement method was then tested in a large multi-group experimental study, where its measurement properties were compared to those of Likert-Type Rating-Scales (LTRSs). The survey included pre-validated psychometric constructs which provided a baseline for comparing the methods, as well as to explore whether certain individual characteristics are linked to respondent choices. Structural equation modelling was used to analyse the survey data. Whilst no strong associations were found between individual characteristics and respondent choices, the results demonstrated that the IRSP is reliable and valid. This study has produced a dynamic measurement instrument that accommodates individual-level differences, not addressed by typical fixed rating-scales.

*This thesis is dedicated to my mother, Rosanna Castaldi;  
my brothers, Jack Chami and Bassam Chami; and,  
Southwold Young Persons Housing Association*

## **Acknowledgements**

Several people were instrumental to the completion of this thesis.

First and foremost, I would like to thank my supervisors Nina Reynolds and James Wallace. Their wisdom and guidance on all matters, both thesis-related and on life's trials, always proved invaluable. I was extremely fortunate to be able to work with, and learn from, two brilliant academics. Not only have I gained two mentors, but also two dear friends.

### **Personal acknowledgements**

Thanks to Ziggy Mwenitete, who was in my life through much of this journey, seeing me through both the ups and the downs. Your words of encouragement and your belief in me during the toughest times will not be forgotten.

Thanks to the other friends I made along the way. Nelarine Cornelius; you gave great advice, and your sense of humour often brightened my day. Desmond Kapofu, Hilary Gunura, and Winifred Huang; some of the loveliest PhD students around. You were there during the tough times when I just needed empathy and a good rant, and you were also there for what would otherwise have been lonely lunches and tea breaks!

Thanks also to my family for being supportive and understanding when I went 'off the radar' for significant periods of time, and for not asking, "So, how's the thesis going?" when it was the last thing I wanted to hear.

### **Professional acknowledgements**

I would like to thank Patrick Oladimeji for doing an excellent job of turning my program specifications into a working reality. The quality of your programming was integral to achieving the finished IRSP.

I would also like to thank the following people for helping me gain access to an ambitious sample frame: William Archer, Gulrez Aktar, Tim Burtwistle and Malcolm Cook. Your assistance was pivotal to my obtaining a large enough sample to test the IRSP.

Last but not least, thanks to Chris Barkby, the Doctoral Student Coordinator, for offering support and a friendly face when operational issues arose. I was always made to feel supported.

# Contents

1.	Introduction .....	1.1
1.1	Background: The Science of Survey Measurement .....	1.1
1.1.1	The Problem with Fixed Rating-Scales.....	1.2
1.1.2	From the Behaviourist to the Cognitive Approach .....	1.4
1.1.3	A Changed Environment: Technology and Survey Research.....	1.5
1.2	A Global Trend: The Rise of Customisation.....	1.6
1.3	Research Gap.....	1.7
1.4	Research Objective.....	1.8
1.5	Research Value.....	1.9
1.6	Research Scope.....	1.10
1.7	Thesis Structure .....	1.10
2.	Literature Review.....	2.1
2.1	Introduction .....	2.1
2.2	Technology and Surveys .....	2.1
2.3	Traditional Survey Questionnaires .....	2.4
2.3.1	Typical Interval Rating-Scale .....	2.5
2.3.2	The Problem of Rating-Scale Length.....	2.6
2.3.3	Numerical labelling.....	2.9
2.3.4	The Existence of the Midpoint.....	2.10
2.3.5	Reliability and Rating-Scale Length .....	2.13
2.3.6	Validity and Rating-Scale Length.....	2.15
2.3.7	Rating-Scale Verbal Labels.....	2.18
2.4	Response Bias.....	2.22
2.4.1	Response Bias and Data Analysis .....	2.26
2.4.2	Existing techniques to measure and correct for response bias.....	2.29
2.5	Cross-Cultural Measurement Equivalence .....	2.33
2.6	Individual Characteristics.....	2.38
2.6.1	Cognitive approach to survey methodology .....	2.40
2.7	Individualised Rating-Scales.....	2.45
2.7.1	The use of computer technology .....	2.50
2.8	Summary .....	2.51
3.	Methodology .....	3.1
3.1	Introduction .....	3.1
3.2	Stance within the Research Domains .....	3.2
3.3	Philosophical Perspective.....	3.5
3.4	Research Design.....	3.7
3.5	Methods.....	3.9
3.5.1	Introduction.....	3.9
3.5.2	Qualitative Phase – Feasibility test and development of the IRSP.....	3.10
3.5.2.1	Sample frame .....	3.12
3.5.2.2	Sample Size.....	3.12
3.5.2.3	Sampling Method.....	3.13
3.5.2.4	Setting .....	3.14
3.5.2.5	Data Capture .....	3.14
3.5.3	Quantitative Phase .....	3.17
3.5.3.1	Selection of a Researcher-Defined Rating-Scale.....	3.18

3.5.3.2	Method .....	3.18
3.5.3.3	Design .....	3.19
3.5.3.4	Validity .....	3.20
3.5.3.5	Sample Size.....	3.22
3.5.3.6	Sample frame .....	3.24
3.5.3.7	Sampling Method.....	3.25
3.5.3.8	Setting .....	3.30
3.5.3.9	Survey items .....	3.31
3.5.3.10	Data Capture .....	3.36
3.5.3.11	Planned Analysis.....	3.36
3.5.4	Ethics.....	3.37
3.6	Summary.....	3.37
4.	The Development of the Individualised Rating-Scale Procedure (IRSP).....	4.1
4.1	Introduction .....	4.1
4.2	Stage 1: Foundations for the Individualised Rating-Scale Procedure (IRSP).....	4.1
4.2.1	Assumptions underlying the development of the ‘rudimentary’ IRSP ...	4.2
4.2.2	Development of the ‘rudimentary’ IRSP.....	4.12
4.2.1.1	Visual aid .....	4.12
4.2.1.2	Instruction order.....	4.15
4.2.3	Interviews: Data Collection and Analysis.....	4.16
4.2.4	Interview Protocol.....	4.17
4.2.5	Round 1 – Interviews 1 and 2 .....	4.19
4.2.5.1	Key findings.....	4.19
Individualised Rating-Scales (IRSs) chosen .....	4.19	
Distinctiveness of response intervals .....	4.20	
4.2.5.2	Key modifications.....	4.21
Verbal anchoring instructions .....	4.21	
Inclusion of Greenleaf’s sixteen items in the IRSP .....	4.23	
4.2.6	Round 2 – Interviews 3-7 .....	4.24
4.2.6.1	Key findings.....	4.25
Individualised Rating-Scales (IRSs) chosen .....	4.25	
IRSP execution times .....	4.27	
The mystery attraction to ‘±10’ .....	4.27	
4.2.6.2	Key modifications.....	4.33
Greenleaf’s Items .....	4.33	
IRSP Instructions .....	4.35	
4.2.6.3	Potential improvements .....	4.47
List of verbal anchors.....	4.47	
An electronic IRSP.....	4.49	
4.2.7	Round 3 – Interview 8 .....	4.53
4.2.7.1	Key finding .....	4.54
Individualised Rating-Scales (IRSs) chosen .....	4.54	
4.2.7.2	Key modifications.....	4.54
The biasing effect of the ‘example’ .....	4.54	
Confusion over the ‘sign’ of disagreement .....	4.58	
4.2.8	Round 4 – Interviews 9-13 .....	4.60
4.2.8.1	Key findings.....	4.60
Individualised Rating-Scales (IRSs) chosen .....	4.60	
IRSP execution times .....	4.60	

The need for IRSP software .....	4.61
List of adverbs.....	4.61
The mystery attraction of ‘±10’ revisited.....	4.66
Personally meaningful IRSs.....	4.70
Greenleaf item nine .....	4.73
Clarity of Instructions .....	4.74
4.2.8.2 Key modification .....	4.75
The inclusion of a bar chart in the electronic IRSP .....	4.75
4.2.8.3 Potential improvement.....	4.76
IRSPv2 .....	4.76
4.2.9 Insights into the Conceptualisation of Agreement and Disagreement...	4.80
4.2.9.1 Bipolar .....	4.80
4.2.9.2 Unipolar .....	4.83
4.2.9.3 Unipolarity, bipolarity and the IRSP .....	4.85
4.3 Stage 2: The IRSP from paper to software .....	4.85
4.3.1 Development of the IRSP Software .....	4.86
4.3.2 The Finished IRSP Software .....	4.87
4.4 Stage 3: Further development of the IRSP .....	4.88
4.4.1 Key Questions .....	4.88
4.4.2 Psychological Measures .....	4.89
4.4.3 Method .....	4.90
4.4.3.1 Concurrent Verbal Protocol-Retrospective Debrief (CVP-RD)	
Interviews .....	4.90
4.4.3.2 CVP-RD Interview Setting .....	4.92
4.4.3.3 CVP-RD Interview Sample .....	4.92
4.4.4 Analysis of Stage 3 .....	4.94
4.4.5 Findings.....	4.95
4.4.5.1 IRSP Survey Software performance .....	4.95
4.4.5.2 IRSPv2 or IRSPv1? .....	4.95
4.4.5.3 Summary of modifications to IRSP instructions .....	4.99
4.4.5.4 Visual Aids .....	4.100
4.4.5.5 Problems reported with items .....	4.104
4.4.5.6 Greenleaf’s items and the IRSP .....	4.105
4.4.5.7 Verbal anchoring and meaningfulness.....	4.106
4.4.5.8 Numerical conceptualisation.....	4.109
4.4.5.9 IRSP vs researcher-defined rating-scales .....	4.111
4.5 Stage 4: Pilot test .....	4.112
4.5.1 IRSP Software Stress Test .....	4.112
4.5.2 Method .....	4.113
4.5.2.1 Online survey .....	4.113
4.5.2.2 Sample .....	4.114
4.5.2.3 Setting and procedure .....	4.114
4.5.3 Findings.....	4.118
4.5.3.1 Sample .....	4.118
4.5.3.2 IRSPv1 vs IRSPv2 .....	4.120
4.5.3.3 Survey software performance .....	4.123
4.5.3.4 Problem items .....	4.125
4.6 Summary .....	4.127
5. Testing the Individualised Rating-Scale Procedure (IRSP).....	5.1

5.1	Introduction .....	5.1
5.2	Stage 1: Establishing Model Fit .....	5.2
5.2.1	Outliers.....	5.3
5.2.2	Testing for normality.....	5.8
5.2.3	Assessing measurement model fit.....	5.14
5.2.3.1	Affective Orientation (AO).....	5.15
5.2.3.2	Personal Need for Structure (PNS).....	5.26
5.2.3.3	Cognitive Style Indicator (CoSI) .....	5.33
5.3	Stage 2: Testing multi-group measurement model equivalence .....	5.39
5.3.1	AO: Measurement Invariance for IRSP vs LTRS groups.....	5.42
5.3.1.1	Factor Structure Equivalence.....	5.42
5.3.1.2	Factor Loading Equivalence and Error Variances Equivalence .....	5.44
5.3.2	PNS: Measurement Invariance for IRSP vs LTRS groups .....	5.45
5.3.2.1	Factor Structure Equivalence.....	5.46
5.3.2.2	Factor Loading Equivalence, Inter-factor Covariance and Error Variances Equivalence .....	5.47
5.3.3	CoSI: Measurement Invariance for IRSP vs LTRS groups .....	5.48
5.3.3.1	Factor Structure Equivalence.....	5.49
5.3.3.2	Factor Loading Equivalence, Inter-factor Covariance and Error Variances Equivalence .....	5.50
5.4	Stage 3: Testing for Validity and Reliability Across Time .....	5.51
5.4.1	Sample Considerations.....	5.52
5.4.2	Test-Retest Reliability.....	5.53
5.4.3	Multitrait-Multimethod (MTMM) Matrix Extending Validity Testing .....	5.59
5.5	Stage 4: Individualised Rating-Scales and Individual Characteristics .....	5.66
5.5.1	Sample.....	5.66
5.5.2	IRS Lengths Chosen.....	5.68
5.5.2.1	Individual Characteristics and IRS Length.....	5.72
5.5.3	IRS Balance Chosen.....	5.75
5.5.3.1	Individual Characteristics and IRS Balance .....	5.79
5.5.4	IRS Verbal Labels Chosen.....	5.85
5.5.5	LTRS versus IRS: Respondent Preferences.....	5.90
5.5.6	Summary .....	5.96
6.	Discussion and Conclusions.....	6.1
6.1	Introduction .....	6.1
6.2	Measurement Model Fit: IRSP vs LTRS .....	6.1
6.2.1	Measurement Model Re-specification .....	6.3
6.3	Testing Multi-Group Measurement Model Equivalence.....	6.5
6.4	Test-Retest Reliability .....	6.6
6.5	Further Test of Validity .....	6.8
6.6	Individualised Rating-Scales (IRSs) and Individual Characteristics.....	6.11
6.6.1	IRSP length .....	6.11
6.6.2	IRS balance .....	6.14
6.6.3	Verbal labels.....	6.16
6.6.4	The Need for a ‘Practice Routine’ .....	6.20
6.6.5	IRSP Feedback .....	6.21
6.6.5.1	Meaningfulness.....	6.22
6.6.5.2	Attention .....	6.23
6.6.5.3	Ease.....	6.23



6.6.5.4	Preference .....	6.24
6.7	Methodological limitations of the study and future research .....	6.24
6.7.1	Nonnormality of the data distribution .....	6.24
6.7.2	Re-specification of the measurement models .....	6.27
6.7.3	Positively worded IRSP feedback items .....	6.29
6.8	Contribution and Implications .....	6.29
6.8.1	Contribution to marketing research methods .....	6.30
6.8.1.1	Cognitive Aspects of Survey Methodology .....	6.30
6.8.1.2	The Imbalanced IRS .....	6.31
6.8.1.3	Construct meaningfulness and response category meaningfulness .....	6.31
6.8.1.4	Stability of IRSs .....	6.33
6.8.1.5	External validity .....	6.34
6.8.2	Contribution to management research practice .....	6.35
6.8.2.1	The preference for IRSP: Increased attention and meaningfulness of categories .....	6.35
6.8.2.2	Response Styles .....	6.35
6.8.2.3	Measuring other concepts .....	6.37
6.8.3	Implications for Business .....	6.39
6.8.3.1	The IRSP and pilot studies .....	6.40
6.8.3.2	The IRSP and cross-cultural studies .....	6.40
6.8.3.3	The IRSP and sample size .....	6.41
6.8.3.4	Construct meaningfulness and rating-scale length .....	6.42
6.8.3.5	Online surveys .....	6.43
6.9	Concluding Remarks .....	6.43
Glossary .....		G.1
References .....		R.1
Appendices .....		On CD attached to thesis

## List of Tables

Table 2. 1 Survey response process based on the cognitive paradigm, (for greater detail refer to Sirken et al., 1999).....	2.43
Table 3. 1 True multi-group experimental design.....	3.20
Table 3. 2 Extraneous variables that affect internal validity.....	3.21
Table 3. 3 Extraneous variables that affect external validity.....	3.22
Table 3. 4 The sample size achieved by each test group. ....	3.23
Table 3. 5 Gatekeepers secured for sample frame .....	3.25
Table 4. 1 Planned order of activity and probes for Interviews.....	4.19
Table 4. 2 Interviewees’ 3-7: Numerical and Verbal Endpoints Chosen.....	4.26
Table 4. 3 Interviewees’ 3-7 Exercise Completion Times.....	4.27
Table 4. 4 Interviewees’ 9-13: Numerical and Verbal Endpoints Chosen.....	4.60
Table 4. 5 Interviewees’ 9-13 Exercise Completion Times.....	4.61
Table 4. 6 Overview of Respondents’ Desire for a List of Adverbs.....	4.62
Table 4. 7 Bipolar view of <i>agreeing</i> and <i>disagreeing</i> : Extracts from interviews.....	4.82
Table 4. 8 Unipolar view of <i>agreeing</i> and <i>disagreeing</i> : Extracts from interviews.....	4.83
Table 4. 9 CVP-RD Interviews: Resulting IRSP modifications made (in order). ....	4.100
Table 4. 10 Protocol-debrief interviews: Problems with items.....	4.105
Table 4. 11 Verbal endpoints chosen by respondents. ....	4.107
Table 4. 12 Numerical Endpoints Inputted by Respondents (Phase Two) .....	4.110
Table 4. 13 Pilot Study with MBA Students: IRSP versions and lab classes. ....	4.116
Table 5. 1 Sample sizes for groups in T1 and T2.....	5.2
Table 5. 2 AO: Unusual cases’ anomaly indices and Mahalanobis scores. ....	5.5
Table 5. 3 PNS: Unusual cases’ anomaly indices and Mahalanobis scores.....	5.6
Table 5. 4 CoSI: Unusual cases’ anomaly indices and Mahalanobis scores.....	5.7
Table 5. 5 AO Assessment of normality (T1 data) .....	5.11
Table 5. 6 CoSI Assessment of normality (T1 data).....	5.11
Table 5. 7 PNS Assessment of normality (T1 data).....	5.12
Table 5. 8 Guidelines for Goodness of Fit Statistics for large samples. ....	5.15
Table 5. 9 AO: Fit Statistics for IRSP and LTRS groups in T1.....	5.17
Table 5. 10 AO: Standardised factor loadings for T1 data. ....	5.19
Table 5. 11 AO: IRSP, extract from the Modification Indices covariances table in AMOS. ....	5.22
Table 5. 12 AO: LTRS, extract from the Modification Indices covariances table in AMOS. ....	5.22
Table 5. 13 AO (minus item 8): Fit Statistics for IRSP and LTRS groups in T1. ....	5.23
Table 5. 14 AO (minus item 8): Standardised factor loadings for T1 data.....	5.23
Table 5. 15 AO (minus item 8): IRSP, extract from the Modification Indices covariances table. ....	5.24
Table 5. 16 AO (minus item 8): LTRS, extract from the Modification Indices covariances table. ....	5.24
Table 5. 17 Original vs Re-specified AO model: Fit Statistics for IRSP and LTRS groups in T1. ....	5.26

Table 5. 18 PNS: Fit Statistics for IRSP and LTRS groups in T1. ....	5.28
Table 5. 19 PNS: Standardised factor loadings for T1 data. ....	5.29
Table 5. 20 PNS Original Model: Factors' Average Variance Extracted and Construct Reliability. ....	5.29
Table 5. 21 Original vs Re-specified PNS model: Fit Statistics for IRSP and LTRS groups in T1. ....	5.32
Table 5. 22 PNS: Factors' AVE and CR for Original Model vs Re-specified Model. ....	5.32
Table 5. 23 CoSI: Fit Statistics for IRSP and LTRS groups in T1. ....	5.34
Table 5. 24 CoSI: Standardised factor loadings for T1 data. ....	5.35
Table 5. 25 CoSI Original Model: Factors' Average Variance Extracted and Construct Reliability. ....	5.35
Table 5. 26 CoSI: Inter-factor correlations T1. ....	5.36
Table 5. 27 Original vs Re-specified CoSI model: Fit Statistics for IRSP and LTRS groups in T1. ....	5.38
Table 5. 28 CoSI: Factors' AVE and CR for Original Model vs Re-specified Model. ....	5.39
Table 5. 29 Stages of multiple group cross-validation, as per Hair et al. (2006). ....	5.40
Table 5. 30 Re-specified AO model: Fit Statistics for IRSP and LTRS groups in T1. ....	5.42
Table 5. 31 Cross-Validation: Equivalence models for multi-group AO in T1. ....	5.45
Table 5. 32 Re-specified PNS model: Fit Statistics for IRSP and LTRS groups in T1. ....	5.45
Table 5. 33 Cross-Validation: Equivalence models for multi-group PNS in T1. ....	5.48
Table 5. 34 Re-specified CoSI model: Fit Statistics for IRSP and LTRS groups in T1. ....	5.49
Table 5. 35 Cross-Validation: Equivalence models for multi-group CoSI in T1. ....	5.51
Table 5. 36 Sample sizes for groups T1 and T2, after the removal of all outliers by factor scores. ....	5.57
Table 5. 37 Spearman correlations between T1 and T2 psychometric factor scores: Test-retest reliability by method (TG1 vs TG4). ....	5.58
Table 5. 38 Spearman correlations between T1 and T2 psychometric factor scores: Test-retest reliability for TG1 (IRSP-IRSP), by IRS change. ....	5.59
Table 5. 39 Campbell and Fiske's MTMM Matrix. ....	5.60
Table 5. 40 Original MTMM Matrix approach. ....	5.62
Table 5. 41 Adapted MTMM Matrix approach. ....	5.62
Table 5. 42 MTMM Matrix: IRSP and LTRS. ....	5.64
Table 5. 43 IRSP Respondents in T1: By national identity and first language. ....	5.67
Table 5. 44 IRS length: Choice to modify, by respondents from T1 to T2. ....	5.71
Table 5. 45 T1 IRSP: Kruskal-Wallis Rankings for IRS Length by Degree Area. ....	5.74
Table 5. 46 T1 IRSP: Kruskal-Wallis Rankings for IRS Balance by Degree Area. ....	5.81
Table 5. 47 T1 IRSP: Mean Respondent IRS Balance Scores Grouped by Degree Area. ....	5.82
Table 5. 48 T1 IRSP: Spearman's rho Correlations between AO and IRS Balance, by Degree Area. ....	5.85
Table 5. 49 T1 IRSP: Verbal Labels Used for Agreement Endpoint. ....	5.87
Table 5. 50 T1 IRSP: Verbal Labels Used for Disagreement Endpoint. ....	5.88

Table 5. 51 T1 IRSP: Verbal Label Symmetry. ....	5.90
Table 5. 52 T1-T2: IRSP-IRSP Test Group, Verbal Label Symmetry. ....	5.90
Table 5. 53 Feedback on the use of the IRSP over LTRS: Test Groups 2 and 3 .....	5.92
Table 5. 54 Test Groups 2 & 3: One-sample t test on IRSP feedback items. ....	5.95
Table 5. 55 Test Groups 2 & 3: One-way ANOVA to test for order-effects on the mean ratings of IRSP feedback items. ....	5.96

## List of Figures

Figure 1. 1 Thesis Structure .....	1.11
Figure 2. 1 Mapping of subjective categories on response categories. ....	2.24
Figure 2. 2 Adapted from the findings of American versus Korean’s use of rating-scales in a study by Riordan and Vandenberg (1994) .....	2.27
Figure 3. 1 Validity Schema as per McGrath and Brinberg (1983). ....	3.3
Figure 3. 2 Adapted from Creswell (2003: 213). ....	3.8
Figure 3. 3 Invitation-to-participate email sent to students at several universities. ....	3.26
Figure 3. 4 Survey welcome page assigning respondents into groups. ....	3.27
Figure 3. 5 Random assignment into test groups by survey entry point. ....	3.29
Figure 4. 1 Example using differing IRSs. ....	4.8
Figure 4. 2 Example using differing IRSs that are aligned .....	4.9
Figure 4. 3 Example using an imbalanced IRS .....	4.10
Figure 4. 4 Example using an imbalanced IRS transformed appropriately .....	4.10
Figure 4. 5 IRSP Instruction Sheet IRSPr1: Interviewees 1 and 2 .....	4.13
Figure 4. 6 Interview Protocol Sheet. ....	4.18
Figure 4. 7 Interviewee 1: IRS Chosen .....	4.19
Figure 4. 8 Interviewee 2: IRS Chosen .....	4.20
Figure 4. 9 IRSP Instruction Sheet IRSPr2: Interviewees 3-7 .....	4.25
Figure 4. 10 Interviewee 5 – Spread of responses .....	4.28
Figure 4. 11 Interviewee 6 – Spread of responses .....	4.31
Figure 4. 12 IRSP Instruction Sheet IRSPr3 Part 1: Interviewee 8 .....	4.52
Figure 4. 13 IRSP Instruction Sheet IRSPr3 Part 2: Interviewee 8 .....	4.53
Figure 4. 14 Interviewee 8: IRS Chosen .....	4.54
Figure 4. 15 IRSP Instruction Sheet IRSPr4 Part 1: Interviewees 9-13 .....	4.59
Figure 4. 16 Interviewee 9 – Spread of responses .....	4.68
Figure 4. 17 Hypothetical example: Mapping one’s ideal IRS onto one with too many intervals .....	4.70
Figure 4. 18 IRSP bar chart visual aid .....	4.76
Figure 4. 19 Agreement/Disagreement continuum represented as a spectrum of shades .....	4.77
Figure 4. 20 IRSP Numerical anchoring instruction represented electronically. ....	4.78
Figure 4. 21 IRSPv2 Numerical anchoring instruction represented electronically. ....	4.79
Figure 4. 22 IRSP2 example .....	4.79
Figure 4. 23 CVP Instructions to CVP-RD Interviewees .....	4.91
Figure 4. 24 CVP-RD Interviews: Age Spread .....	4.93
Figure 4. 25 CVP-RD Interviews: Gender Spread .....	4.93
Figure 4. 26 CVP-RD Interviews: First Language Spread .....	4.93
Figure 4. 27 Respondents’ chosen number of categories on versions 1 and 2 of the IRSP .....	4.96
Figure 4. 28 CVP-RD Interviewee 9: Spread of responses to Greenleaf items .....	4.96
Figure 4. 29 CVP-RD Interviewee 15: Spread of responses to Greenleaf items .....	4.97
Figure 4. 30 CVP-RD Interviewee 3: Spread of responses for Greenleaf items. ....	4.102

Figure 4. 31 CVP-RD Interviewee 3: Spread of responses for main survey items....	4.102
Figure 4. 32 IRSPv1 CVP-RD Interviewee 4: Spread of responses for Greenleaf’s items. .....	4.104
Figure 4. 33 IRSPv1 CVP-RD Interviewee 4: Spread of responses for main survey items. ....	4.104
Figure 4. 34 MBA Pilot: Age spread .....	4.118
Figure 4. 35 MBA Pilot IRSPv1: Respondents’ ethnicities .....	4.119
Figure 4. 36 MBA Pilot IRSPv2: Respondents’ ethnicities .....	4.119
Figure 4. 37 MBA Pilot IRSPv1: Respondents’ first language. ....	4.120
Figure 4. 38 MBA Pilot IRSPv2: Respondents’ first language. ....	4.120
Figure 4. 39 MBA Pilot: Histograms showing no. categories chosen by IRSP group. .....	4.121
Figure 4. 40 MBA Pilot: Histograms showing index of dispersion by IRSP group. .	4.122
Figure 4. 41 Problem-words flagged by MBA students in their feedback sheets.....	4.126
Figure 4. 42 Items to which the problem-words belong. ....	4.127
Figure 5. 1 Big Five Index 10-item.....	5.3
Figure 5. 2 Scatterplot AO: Illustrating outliers through respondents’ anomaly indices. .....	5.5
Figure 5. 3 Scatterplot PNS: Illustrating outliers through respondents’ anomaly indices. .....	5.6
Figure 5. 4 Scatterplot CoSI: Illustrating outliers through respondents’ anomaly indices. .....	5.7
Figure 5. 5 AO Original Measurement Model. ....	5.16
Figure 5. 6 Re-specified AO model: Standardised regression weights shown for both groups in T1. ....	5.25
Figure 5. 7 PNS Original Measurement Model. ....	5.27
Figure 5. 8 Re-specified PNS model: Standardised regression weights shown for both groups in T1 .....	5.31
Figure 5. 9 CoSI Original Measurement Model.....	5.33
Figure 5. 10 Re-specified CoSI model: Standardised regression weights shown for both groups in T1 .....	5.37
Figure 5. 11 AO Re-specified model: Multi-group factor structure equivalence (TF model) parameter estimates.....	5.44
Figure 5. 12 PNS Re-specified model: Multi-group factor structure equivalence (TF model) parameter estimates.....	5.47
Figure 5. 13 CoSI Re-specified model: Multi-group factor structure equivalence (TF model) parameter estimates.....	5.50
Figure 5. 14 Time elapsed between T1 and T2 for all respondents. ....	5.53
Figure 5. 15 AO Repeat-measures structural equation model. ....	5.55
Figure 5. 16 CoSI Repeat-measures structural equation model.....	5.56
Figure 5. 17 T1 IRS Lengths Used.....	5.69
Figure 5. 18 T1 IRS Balance Used.....	5.77
Figure 5. 19 T1 IRSP: Distribution of IRS Balance Scores for Respondents within the Business Discipline. ....	5.82

Figure 5. 20 T1 IRSP: Distribution of IRS Balance Scores for Respondents within the Physical Sciences Discipline.....	5.83
Figure 5. 21 IRSP feedback questions posed to TG2 and TG3 at end of survey.....	5.91
Figure 5. 22 Test Groups 2 & 3: Spread of Responses to ‘Attention’ Item.....	5.92
Figure 5. 23 Test Groups 2 & 3: Spread of Responses to ‘Meaningful’ Item.....	5.93
Figure 5. 24 Test Groups 2 & 3: Spread of Responses to ‘Preference’ Item.....	5.93
Figure 5. 25 Test Groups 2 & 3: Spread of Responses to ‘Ease’ Item.....	5.94
Figure 6. 1 Patterns of correlations identified through the MTMM matrix.....	6.9

# **Chapter 1. Introduction**

---



# 1. Introduction

## 1.1 Background: The Science of Survey Measurement

The evolution of measurement science has benefited a number of fields, including business, politics, and medicine. *Measurement* was defined as a procedure for assigning numerals to objects, events, or persons according to a rule (Campbell, 1928). In particular, *measurement* referred to the assignment of numerals in such a way as to correspond to *different degrees of quality/property* of some object, event or person (Duncan, 1984) such as a person's attitude toward a subject on an attitudinal continuum with fixed numerical values. In a broader sense, *measurement* is part conceptual and part empirical (Bagozzi, 1984).

*“Measurements achieve meaning in relation to particular theoretical concepts embedded in a larger network of concepts, where the entire network is used to achieve a purpose. The purpose may be to solve problems and answer research questions for [...] researchers, brand managers and advertising account executives. More broadly, the purpose is to obtain understanding, explanation, prediction, or control of some phenomenon.”*

(Bagozzi, 1994: 2)

There was a spectacular growth in *measurement* through survey research during the twentieth century (Sirken and Schecter, 1999). This was fuelled by the ever-growing need for businesses to measure a plethora of phenomena, such as a changing market environment, consumer attitudes, and the human resources landscape. Additionally, globalisation drove this need for businesses to maintain a “finger on the pulse” in order to remain responsive and competitive. Survey research has provided a means through which

businesses have been able to quantitatively *measure* the concepts of interest to them. Whether that is to find out what their consumers think of their new product line, or to better understand how to increase staff retention rates, it has been a key information gathering approach to help them achieve their business objectives. Even beyond business, survey questionnaires are the dominant data collection method in psychology and the social sciences in general (Rohrmann, 2003). The most widely used response mode in survey research is, arguably, the fixed interval rating-scale<sup>1</sup> (Bagozzi, 1994, Rohrmann, 2003). Typically, survey questions and survey response formats have been standardised and fixed within a survey, across respondents (Herrmann, 1999). This has resulted in problems in measurement.

### **1.1.1 The Problem with Fixed Rating-Scales**

Measurement problems have been one of the main obstacles to the advancement of social science research (Blalock, 1979). Bollen and Barb (1981) pointed out that one manifestation of this problem is the measurement imprecision resulting when continuous concepts are measured on rating-scales containing relatively few categories. The debate surrounding what the optimal number of response alternatives are for a rating-scale has received much attention with conflicting recommendations over the last ninety years (Boyce, 1915, Conklin, 1923, Symonds, 1924, Champney and Marshall, 1939, Cronbach, 1950, Jacoby and Mattel, 1971, Miller, 1956, Alwin, 1992, Weng, 2004).

---

<sup>1</sup> The word 'scale' has been used to mean several things in the literature. It is sometimes used when referring to the measurement tool that respondents use to rate their opinions, and it is also used when referring to a list of items that measure an overall latent variable. To avoid confusion, throughout this thesis the term 'rating-scale' has been used when referring to the measurement tool used by respondents, and 'scale' is used when referring to a pre-validated list of items that measure a latent variable.

The essential question from a measurement point of view is whether, for any particular survey question, there is an optimal number of response categories beyond which there are no further improvements in discrimination along an attitudinal continuum (Garner, 1960, Cox III, 1980, Komorita and Graham, 1965, Miller, 1956). If too few alternatives are used, the rating-scale is too coarse, resulting in a loss of the discriminative powers of which respondents are capable. However if a rating-scale is graded too finely, it would be beyond the respondents' limited powers of discrimination and the data is contaminated with error (Guilford, 1954). This is reiterated by Cox's (1980: 408) precise definition:

*“A scale with the optimal number of response alternatives is refined enough to be capable of transmitting most of the information available from respondents, without being so refined that it simply encourages response error.”*

The problem is complicated further when other attributes (of fixed rating-scales) found to impact on data quality are considered; for example, the choice of verbal and numerical labels, and the existence of a midpoint. Whilst there are existing techniques for removing the resultant data error, they are few and have received criticism due to the fact they are post-hoc and may require sophisticated statistical application. These issues are further explored in the Literature Review.

The next two sections explore why these problems exist and how technology can enable some of these issues to be addressed.

### 1.1.2 From the Behaviourist to the Cognitive Approach

Historically, survey questions and survey response formats (e.g. rating-scale length) have been standardised and fixed within a survey, across respondents. This is unsurprising given that, for decades, questionnaire design had been guided by behaviourism which “viewed questions as stimuli that elicited responses much in the same way as a bell elicited saliva from Pavlov’s dog” (Herrmann, 1999: 267). Although some survey methodologists recognised that the response process was a complicated function of questions (Payne, 1951), question answering was seen fundamentally as a stimulus-response process. After five or so decades of extensive use, the behaviourist concept of stimulus and response to questionnaire design had reached the limits of its usefulness, with a cognitive approach becoming more popular (Herrmann, 1999). For example, in the early 1980’s leading survey methodologists and cognitive psychologists came together at an Advanced Research Seminar on Cognitive Aspects of Survey Methodology (CASM I Seminar) to determine whether cognitive psychology might provide a better understanding of how respondents answer survey questions (June 1983). In contrast to the behaviourist approach, these experts concluded that answers to questions were *derived*, not *elicited*, through a series of cognitive processes: perception, encoding, comprehension, memory retrieval, thought, editing of potential answers, and expression of the answer (Jobe and Mingay, 1991). Herrmann (1999) points out that this cognitive approach enabled a principled explanation of measurement error, where response error was a function of the kinds of ideas embraced by a question and the knowledge and cognitive skills that a respondent possessed. This way of looking at survey measurement highlights the fundamental flaw in fixed rating-scales. They are unable to adjust for individual-level variations in the interpretation and use of these

response formats. As such, measurement error might be reduced if researchers were to take a more cognitive approach to survey design.

### **1.1.3 A Changed Environment: Technology and Survey Research**

Whilst survey research was originally influenced by the behaviourist approach, technology (or lack of) undoubtedly placed limitations on what could be achieved in terms of questionnaire development. Questionnaires were mostly paper-based, and encompassed a standardised list of questions and fixed response formats. However, the introduction of the Internet, developments in software applications, and the shift to a cognitive approach to survey research, has fuelled a wave of online research activity.

Initially, migration efforts to online market research focused on research activities such as: concept and product testing; advertising and brand tracking; customer satisfaction measurement; usage and attitude studies; opinion polling; and, qualitative research (Miller, 2006, Comley, 2007). However, the opportunities available through online research have already given rise to a wave of new research activities and ever-developing survey methods.

*“Today’s online researchers are often not just interested in migrating traditional research methods to the online medium. Instead, they are looking to take advantage of the interactive nature afforded by the online environment to conduct studies that might have been difficult, if not impossible, to conduct in the offline environment.”* (Miller, 2006: 110)

Online surveys have already been shown to be a practical and valuable resource for social scientists. As such, it is appropriate that further contributions are made to developing this research method. With more than 80% of online research spending in the US being devoted to researching the *attitudes* and *activities* of consumers (Business to Consumer) (*Inside Research* 2006), these areas are worthy of particular interest when making future developments to online survey methods. The online survey method is a tool that is very well positioned to support the aims of a cognitive approach to survey research, and may also aid in the effort to reduce measurement error resulting from fixed rating-scales.

## **1.2 A Global Trend: The Rise of Customisation**

The focus on a measurement problem which stems from a standardised approach to measurement, by proposing it be addressed through a more dynamic approach, is also in line with a more general trend in the marketing field; marketing, as a whole, has been shifting from a mass marketing approach to a more customised one, fuelled especially by the advent of technology such as the Internet (Poynter, 2007). Products and services have for some time been moving further and further towards customisation, and the Internet has provided opportunities for businesses to ‘get to know’ their customers better (through processes such as Customer Relationship Management), and tailor their offerings to match these differing consumer requirements. If one were to look at traditional survey questionnaires as a ‘standardised’ approach to measurement (across respondents) whereby rating-scales are fixed, technology has now afforded us the opportunity to ‘customise’ our approach to measurement for every respondent.

*“Survey designers can help participants perform their tasks more accurately by developing plans that can be adapted to the participants’ knowledge, beliefs, and behaviors much like user models are designed to promote more accurate human-computer interaction.”* (Conrad, 1999: 308)

There is a body of evidence that suggests that a more individualised approach to measurement would greatly improve accuracy of data capture. This is outlined in the Literature Review together with the methodological justification for taking this new approach to measurement. This is in keeping with the general shift from ‘standardised’ to ‘customised’; the current trend in marketing and industry more generally. At a time when businesses are able to provide consumers with customised products and services, researchers should also be able to adapt to the individual characteristics of research participants in order to improve measurement, by customising the methods used.

### **1.3 Research Gap**

It is critical for researchers to be confident of the quality of survey data. The need to improve the methods for capturing data, and as such minimising measurement error, has always been high on the agenda in the measurement sciences, and still is. Problems with data quality can be traced back to measurement method *design*; the choices made by the researcher in their creation of a standardised measurement instrument. This has been shown to affect the way respondents interpret and respond to these instruments (rating-scales), resulting in measurement error (Hui and Triandis, 1989). Current methods for removing measurement error are post-hoc (i.e. not preventative) and have been shown to be problematic. Now that fewer scientists think of survey research through a simplistic

behavioural approach, there is scope to innovate further through the creation of measurement methods that take respondents' individual cognitions into consideration. Add to this the technological capability and the Internet, not available to researchers in the past, and we now have opportunities to create superior methods of measurement that may further enhance data quality.

#### **1.4 Research Objective**

The overall objective of this research is to address a methodological problem; the reduction of measurement error from data obtained through survey research. Specifically, this thesis aims:

*To develop and test a measurement instrument capable of having respondents individualise their own rating-scales for use in online surveys.*

This objective naturally breaks down into two parts:

1. The development of a measurement instrument: Achieved through taking a qualitative approach which allows for the exploration of possible solutions.
2. The testing of the measurement instrument: Achieved through taking a quantitative approach which provides a test environment designed to demonstrate the instrument's validity and reliability.

As such, this research adopted a mixed methodology to satisfy the overall objective.



## 1.5 Research Value

The use of technology to develop an electronic method enabling respondents to individualise rating-scales offers a novel way for researchers to engage respondents in online surveys. This could be particularly useful in an environment where response rates are falling, and respondents have become disinterested as a result of “seeing the same old same old”. A novel way of interacting with respondents through data capture would increase involvement in the process and possibly response rates. In this context respondents would also think more carefully about their responses, therefore increasing the accuracy of self reports and enhancing data quality.

Another advantage of developing an electronic method for having respondents individualise a rating-scale relates to the reduced impact of response styles in the online environment.

*“The potential to eliminate or reduce certain biases, such as acquiescence, extreme responding, and social desirability, has also fuelled the adoption of online research methods. Many online research practitioners value the anonymous nature of the Internet environment, where survey respondents can be free to express allegedly truthful attitudes and opinions without the unwanted influence of interviewers in survey administration.”* (Miller, 2006: 112)

An electronic individualised rating-scale provides the opportunity for respondents to use a more meaningful rating-scale in an online setting, which could help reduce response bias.

This research could contribute to the CASM (Cognitive Aspects of Survey Methodology) movement. A deeper understanding of how respondents answer survey questions is needed. A measurement method that factors in the cognitive traits of respondents is in line with the type of research that contributes to furthering the CASM movement.

## **1.6 Research Scope**

Whilst individualised rating-scales have obvious applicability in a cross-cultural research context, the scope of this research is to develop the method itself and within a single culture. The mono-cultural validation can be tested cross-culturally in subsequent research.

## **1.7 Thesis Structure**

The rest of this thesis is divided into five chapters (shown in Figure 1. 1), and can be divided into four main areas:

1. The theoretical framework and methodology of the research (Chapters 2 and 3);
2. The qualitative analysis and findings (Chapter 4);
3. The quantitative analysis and findings (Chapter 5);
4. The interpretation of the results and their implications (Chapter 6).

More specifically, the Literature Review Chapter (Chapter 2) will expand on some of the issues raised here, such as the shift to a more cognitive approach to survey research, growth in online research activities, and measurement problems symptomatic of traditional survey design. One key area covered is how researchers' measurement choices regarding rating-scale length have been linked with the introduction of error in responses, in particular

through the manifestation of response bias. The detrimental impact of response bias on data quality is outlined, along with the current popular techniques for correction, and their drawbacks. The impact of inappropriate standardised measurement instruments, and subsequent measurement error, is discussed in the cross-cultural research context where this proves even more problematic. An argument for improving data quality through measurement design is presented, along with the proposed way this might be achieved, namely, by using individualised rating-scales.

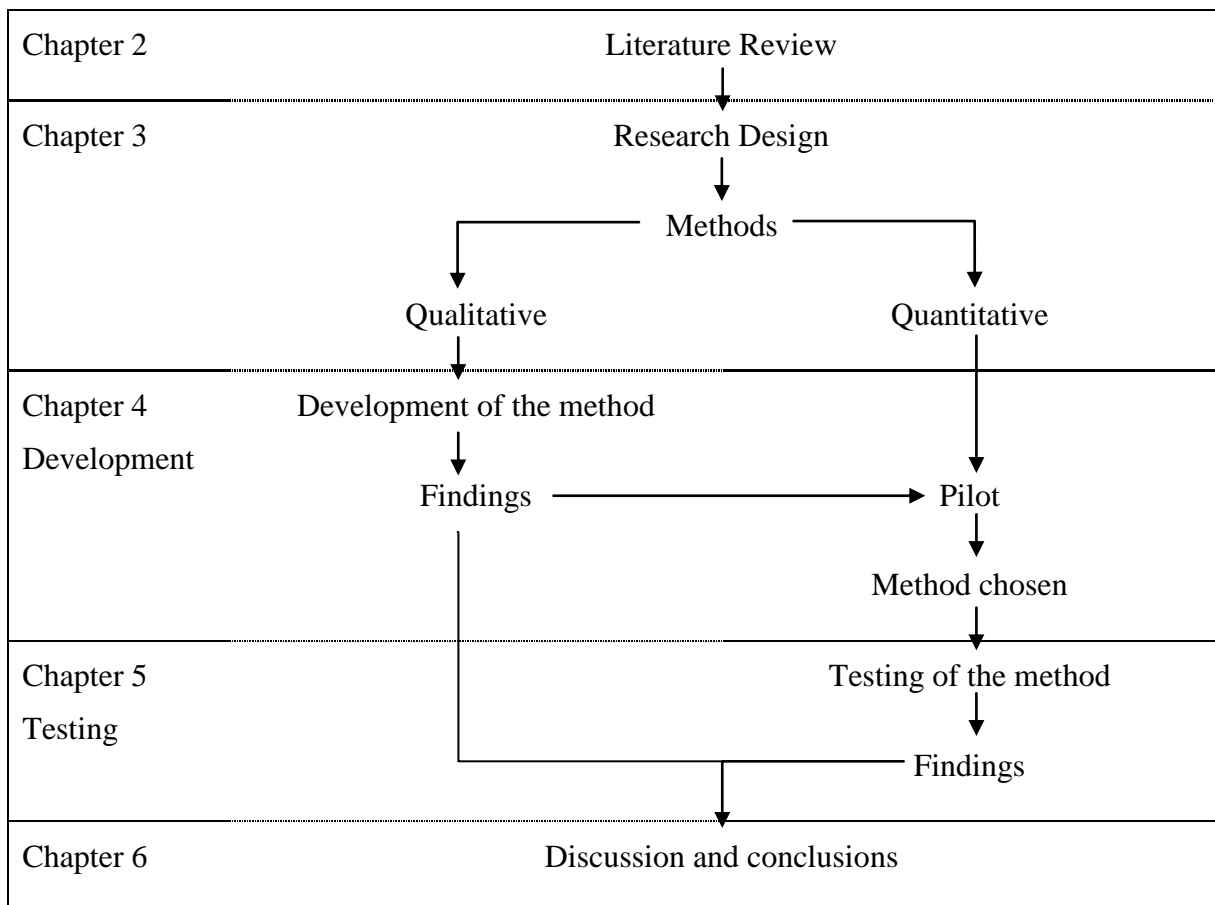


Figure 1. 1 Thesis Structure

Chapter 3 describes the foundations of this research, the philosophical stance, the proposed research design, and the methods chosen. The section on methods is divided into two main areas; one which addresses the first part of the research objective which seeks to ‘develop’ a measurement instrument through qualitative exploration, and the second which aims to ‘test’ it through a robust quantitative approach. Outlined in each are the justifications for the sampling approach employed, analysis, and other substantive methodological choices.

Chapter 4 presents the four key stages that were used to develop the measurement instrument. Stages 1 and 3 outline how the qualitative interviews led directly to development of the method, and provided the underlying insights about respondent cognitions. Stage 2 details the steps taken to transform the measurement instrument to an electronic tool, through software design. Stage 4 describes the quantitative pilot test conducted, and final adjustments made to refine the measurement instrument.

Chapter 5 presents the quantitative *testing* of the measurement instrument. It compares three psychometric scales on the new measurement instrument against the same psychometric scales on a pre-existing, popular measurement instrument, through a multi-group experimental design. The chapter is divided into four stages. In Stage 1, confirmatory factor analysis (CFA) of the data was undertaken to establish loose cross-validation across two groups (pre-existing versus new measurement instruments) for three measurement models. Stage 2 applies further cross-validation testing across both groups. This was done through CFA analyses to test for measurement equivalence, using increasingly tighter cross-validation procedures. Stage 3 assesses test-retest reliability for both groups. Stage 4

examines respondent measurement choices and whether they are related to their individual characteristics.

The final chapter (Chapter 6) presents a discussion of the research findings and the implications for researchers. The limitations of this study are highlighted, along with suggestions for future research to address these limitations. Additional areas of further research considered to be worthy of interest are outlined; due to their likely contribution to the field of survey measurement and the CASM movement.

## **Chapter 2. Literature Review**

## **2. Literature Review**

### **2.1 Introduction**

This chapter begins by setting the context through which researchers can now conduct research activities, and highlighting the opportunities afforded by the advent of the Internet and technological improvements. It is in this changed external environment that measurement problems can now be examined through a new lens. Specifically, the literature review will discuss:

- The traditional survey questionnaire and the measurement instrument choices taken by researchers, along with their impact on data quality;
- The manifestation, in particular, of one type of measurement error which adversely affects data quality; response bias;
- The cross-cultural research context, and how these measurement issues are exaggerated in this setting;
- The cognitive approach to survey methodology, and a closer examination of how individuals' characteristics can be accounted for through the measurement process;
- The assertion that rating-scales can be individualised and improve data quality, along with a discussion on how this might be achieved.

### **2.2 Technology and Surveys**

Survey research has always been limited by the technology available for its employment. In the early days, it was a paper-based process, with questionnaires being posted out to addresses. In America during the early 1970's, telephone technology provided a new medium to conduct surveys, and it quickly became a major survey mode

due to its relative low cost and broad reach, given it had almost 90% household coverage rates (Dillman and Smyth, 2007). It was also increasingly effective in the application of random sampling methods. However, today it is argued that researchers are now in the beginning stages of another major change in survey methods. Cultural changes in how the telephone is viewed and used have led to reductions in the advantages that previously drew researchers to the telephone-administered questionnaire (i.e. coverage and cooperation rates), and technological innovation has produced a new mode of data collection, the online survey (Dillman and Smyth, 2007). Online market research, has been defined as all research that is conducted in some way using the Internet, and typically takes the form of online surveys on web pages (Comley, 2007). There is no doubt that researchers around the world have embraced the significant opportunities for market research afforded by the Internet. According to *Inside Research* (2006), an industry newsletter that surveys firms on their expenditure on online research activities, a third of all market research (by value) in the US was conducted online in 2006. In other countries, such as the Netherlands, adoption has been even greater with more than half of all research being carried out online by 2005. This shift to online research has been a trend around the world, with online research spending in Europe going from 77 million Euros in 2001 to 288 million Euros in 2006 (Inside Research, 2006). This also shows that whilst the US represents approximately 80% of worldwide online market research, European and Asian adoption has gained momentum with current growth rates outpacing those in the US.

Comley (2007) summarises some of the key predictions made about the future of online research:



- Online research will increase: It is very likely that we will see online research accounting for at least 30% of all research worldwide by 2010.
- Online costs will decrease in the short term: Competition between growing online panels will force prices down.
- Online response rates will continue to decline: The decline in response rates is largely a reflection of societal changes, and online research is not immune from this overall decline which has affected the more traditional research modes for some time now.

Given Comley's (2007) above predictions, it is important for researchers to innovate, as keeping abreast of developments in online survey research is critical to remain competitive and current. The Internet offers a much cheaper alternative to other methods (certainly in the short-term), and its reach is global. Online surveys will provide researchers with an ever increasing amount of flexibility, as internet penetration increases and technology improves. This new platform has already been embraced by many research agencies around the world, and is set to continue. Researchers might also wish to consider new and innovative ways of engaging respondents in online surveys, given response rates online are set to continue falling. This is an area that will need further advancement by the sector.

Whilst some of the key advantages of online research are obvious (e.g. speed, low cost, versatility, data quality) and have been extensively explored in the literature (e.g. see Miller, 2006, Comley, 2007, Sills and Song, 2002), the concern that has received the most attention has been with regards to sampling. This concern is about the potential for 'coverage errors' to occur in online survey research, where some population members

are excluded from a drawn sample because they do not have Internet access. This concern is becoming less of an issue with more and more people becoming connected. The issue of non-connected respondents has been found to vary quite significantly across demographic groups. For example, in the US, Internet penetration in 2004 was only 15% among individuals who did not graduate from high school. Yet 85% of college graduates were online (Miller, 2006). Over time, however, Internet access differences across groups will dissipate (Sills and Song, 2002), and the degree to which coverage will impact upon representativeness depends entirely on the requirements of the sample for individual projects. An important point however is that all methods (i.e. telephone-administered, online surveys etc) have the potential for large non-sampling errors to occur whether relating to coverage, non-response, or measurement, all survey administration methods have advantages and disadvantages. The design flexibility, geographic reach, anonymity, and minimised interviewer error, of Internet surveys have been shown to be superior to telephone and mail delivery methods (Sills and Song, 2002). For populations that are connected and technologically savvy, the cost, ease, speed of delivery and response, ease of data cleaning and analysis, all weigh in favour of the Internet as an administration method for survey research.

Given online surveys have already been shown to be a practical and valuable resource for social scientists, it is appropriate that further contributions be made to developing this research method.

### **2.3 Traditional Survey Questionnaires**

The issue of measurement with traditional survey questionnaires has been a consideration for some time. Issues such as ‘the art of asking questions’ (Payne, 1951,

Rugg and Cantril, 1944, Schaeffer and Presser, 2003), appropriate choice of response formats (Symonds, 1924, Bardo and Yeager, 1982, Conklin, 1923, Likert, 1932, Ferguson, 1941), and scale development for the measurement of latent variables (Boyce, 1915, Werner, 1948), have been investigated for decades. Over the years, the literature has yielded a plethora of varied conclusions and recommendations on these key issues of questionnaire design. Discussed next, are the traditional considerations applied by researchers to questionnaire design, in particular to measurement methods, together with some of the varied recommendations on those issues.

### **2.3.1 Typical Interval Rating-Scale**

In the typical interval rating-scale (hereafter referred to as rating-scale), numbers are assigned along a continuum to indicate differences in the degree of a property, such that the differences from number to number are equal across the range of the rating-scale (Bagozzi, 1994). Different formats have been used to create scales that are interval in character. One of the most frequently used is the Likert-type rating-scale (also referred to as the summated rating-scale). It accompanies a series of statements regarding an attitudinal object for which a respondent evaluates agreement or disagreement (for a more detailed outline of the Likert scale see Likert (1932)). Typically, the rating-scale would consist of five or seven steps with 'strongly agree' and 'strongly disagree' at either endpoint, and verbal labels clarifying the degree of agreement/disagreement along the middle-intervals. Likert-type rating-scales are often used to measure attitudes towards objects, brands, people, products, services, and so frequently appear in survey questionnaires as the chosen response mode. The semantic differential is another popular choice among researchers, and typically consists of seven-point bipolar scales anchored with adjectives at either end. Although seven-point rating-scale steps are the

most common, five-point, nine-point and eleven-point rating-scales are sometimes used (Bagozzi, 1994). The Likert-type and the semantic differential are the most commonly used rating-scales.

### **2.3.2 The Problem of Rating-Scale Length**

Early last century, researchers began reviewing the number of alternatives employed in rating-scales (Boyce, 1915). However, around this period, there was little research into systematically examining the effects of the number of response categories. Conklin (1923) raised this issue and also recommended the use of nine-point rating-scales over thirteen (for bipolar scales) with untrained respondents, due to some of the options being neglected when the more refined (longer) rating-scale was employed (i.e. scale attenuation). Symonds (1924: 79) however, was first to stress that the problem was primarily one of reliability (inter-rater correlation), suggesting that “a seven-level scale was optimal for rating personality traits.” He argued that fewer steps should be used if the trait was obscure, if the respondents were untrained and only moderately interested, or if a number of ratings of different aspects of the object rated were to be combined (Symonds 1924). Champney and Marshall (1939) later demonstrated that when the respondent was trained and interested, the optimal number of alternatives may be as many as twenty-one. Their reasoning was that extra information can be obtained from the respondents (using more finely graded rating-scales). Whilst this might be true, there is now the argument that there is a greater amount of measurement error present in the data when rating-scales with too many categories are used (Cox III, 1980, Preston and Colman, 2000). Cox III (1980: 409) argued that,

*“as the number of response alternatives is increased beyond some minimum, the demands placed upon a respondent become sufficiently*

*burdensome that an increasing number of discrepancies occur between the true scale value and the value reported by the respondent. Thus, though the information transmission capacity of a scale is improved by increasing the number of response alternatives, response error seems to increase concomitantly."*

As such, rating-scale lengths need to find that balance between the benefit of information gleaned from true variance within the data, and the cost of the data being clouded by error variance. This is complicated further when we consider that this balance potentially differs by respondent.

Miller's (1956) seminal work argued the merits of the seven-category rating-scale (referring to it as the magical number seven) and the diminished utility of response formats with more than seven response categories. He argued that the "span of absolute judgment and the span of immediate memory impose severe limitations" (1956: 95) on the ability for people to work with lengthy sets of response categories. Miller also pointed out that psychologists had been using seven-point rating-scales for a long time in the belief that finer categories would not add much to the usefulness of the ratings.

While such findings have been applied to the question of the optimum number of response categories, several problems exist. As follows, the most obvious is that survey measurements of attitudes/objects and other subjective responses are not comparable to objective stimulus-centred response scales, the substantive focus in Miller's (1956) study. Alwin (1992) points out that they essentially involve assessments of internal states rather than perceptions of external physical absolutes. Furthermore, according to social judgement theory, variables like attitudes are best thought of in terms of

“latitudes” or “regions” on a rating-scale and not by single points on a latent continuum (Sherif et al., 1965). In addition, response categories used as rating-scale points in survey research are often given explicit verbal and numeric labels, and therefore the meaning of these categories may be enhanced from the perspective of the information conveyed (Sarvis, 1988). The researcher should be cautious and think more specifically about the degree to which Miller’s conclusions can be applied to a specific research situation.

There have, however, been many theorists that have supported the conclusions made by Miller based on the results of independent studies. Finn (1972) for example, found that the optimal number of response categories was seven taking into account the reliability of ratings and the desire to maximise variances of ratings (information-giving factor). In addition, Lehmann and Hulbert (1972) concluded that if the focus of a study is on individual behaviour, or if individual scales are to be analysed, five- to seven-point rating-scales should be used. On the flip side, if the researcher is interested in averages across people, or will average or aggregate several individual items, then they recommend that two or three rating-scale points are generally good enough.

Over the decades, the conflicting recommendations for number of response categories continued, with some joining camps that argue that as few as two or three response alternatives are appropriate (Peabody, 1962, Jacoby and Mattel, 1971), and others arguing that this number is generally incapable of transmitting very much information and has the additional disadvantage of frustrating some respondents (Cox III, 1980). Some suggested that the optimal number of response categories be as high as twenty-five, and that information is not lost by increasing the number of rating categories

(Garner, 1960, Guilford, 1954). However, the ability of respondents to make fine distinctions between adjacent response categories on a long rating-scale has been questioned (Green and Rao, 1970), as has the meaningfulness of the answers given (Viswanathan et al., 2004).

### **2.3.3 Numerical labelling**

As discussed above, when respondents are presented with fixed rating-scales (such as Likert or semantic differential rating-scales) the researcher imposes on the respondent his/her vision of how the concept should be conceptualised, and selects the appropriate rating-scale accordingly. The researcher typically decides how many intervals to include and what numerical and verbal labels should be attached. However, respondents have been shown to use the numeric values provided on a rating-scale to disambiguate the meaning of rating-scale labels (Schwarz et al., 1991a, Schwarz et al., 1991b). Schwarz et al. (1991a) found that if numeric values range from 0 to 10, their very structure suggests that the researcher is interested in the absence or presence of the attribute to which the rating-scale pertains. They found that if the numeric values range from - 5 to +5, including a zero at the midpoint, their structure seems to suggest that the absence of the attribute corresponds to zero, whereas the negative values refer to the presence of its opposite. In summary, Schwarz et al. (1991a) found that rating-scales that provide a continuum from negative to positive values (e.g. -3 to 3) may indicate that the researcher has a bipolar conceptualisation of the respective dimension, whereas rating-scales that present only positive values may indicate a unipolar conceptualisation. The choice of numeric values either facilitates or dilutes the polarity implications of the endpoint labels that are provided to respondents. As a result, researchers have been

advised to attempt to match the numeric values that they provide to respondents with the intended conceptualisation of the underlying dimension as uni- or bipolar.

#### **2.3.4 The Existence of the Midpoint**

Whether an even or an odd number of response categories are chosen will depend on whether the researcher requires a neutral point on the rating-scale or not. Researchers have long known that when a middle response is offered it will be chosen by more respondents than will volunteer that answer when it is not offered (Presser and Schuman, 1980). Some researchers favour offering a middle response on the grounds that if it is not explicitly offered respondents with neutral views may feel forced to give false responses; others prefer to exclude it in order to persuade respondents to make a clear choice (Kalton et al., 1980).

Presser and Schuman (1980) raise three important issues that researchers may want to consider before deciding whether to include or omit a middle response. Firstly, they affirm that when survey investigators decide against offering a middle response, they are usually assuming that a middle category consists largely of responses from those who lean toward one or the other polar alternatives, though perhaps with little intensity. “Thus it is legitimate to press respondents to choose one of these alternatives, rather than allowing them to take refuge in a middle position” (Presser and Schuman, 1980: 71). Secondly, some researchers may want to omit the middle category, if they believe it may attract those who have no opinion on the issue and would rather adopt a noncommittal position than say ‘I don’t know’. Thirdly, researchers may want to include the middle position if they believe that respondents who opt for it genuinely do hold a neutral view, and if forced to choose a polar alternative this will contribute some



form of random or systematic error. This is supported by O'Muircheartaigh et al. (1999) who conclude that offering a middle alternative in rating-scales reduces the amount of random measurement error and does not affect validity. The investigator may also want to consider a point raised by Guy and Norvell (1977), who point out that when the neutral response is omitted, there is a greater tendency for raters to give no response (non-response bias).

Several split-ballot experiments have been conducted to compare distributions of responses obtained when a middle position is offered and when it is not (Rugg and Cantril, 1944, Stember and Hyman, 1949, Schuman and Presser, 1977, Kalton et al., 1978, Presser and Schuman, 1980). What predominates from these experiments is the increase in the proportion of respondents endorsing the neutral view when a middle response is included. This increase has varied from as little as 3% for some questions to 20% or more for others. Presser and Schuman (1980) point out that one way of interpreting this increase is that respondents make different assumptions about the information being requested by the two question forms; when the middle response is omitted, they tend to assume that they are meant to decide which way they lean, and hence feel constrained to avoid the middle ground. This interpretation supports the conclusions drawn by Payne (1951). He recommends that if the intent of the question is to discover more definite *convictions* the middle option should be offered, but that if the question aims to find out respondents' *leanings* it should not.

With regard to the effect which including or omitting a middle response would have on the results of a study and the conclusions formed, there are conflicting views. Presser and Schuman (1980) concluded from their study that when the middle alternative is

offered, almost all the change to the middle position comes from a proportionate decline in the polar positions. They state that rating-scale format “is usually unrelated to the univariate distribution of opinion once middle responses are excluded from analysis,” (Presser and Schuman, 1980: 83). However, Kalton et al. (1980: 77) assert that “to assume that in general the proportionate split between polar positions will be unaffected by whether or not a middle option is offered would be dangerous.” In fact this proportionate decline in polar positions failed to replicate itself in two of the three split-ballot tests in their study. Guy and Norvell (1977) reported that composite scores on Likert-type rating-scales were significantly affected by the omission of the neutral response. Bishop (1987: 229) adds to this argument by affirming that offering respondents a middle alternative “will generally make a significant difference in the conclusions that would be drawn about the distribution of public opinion on an issue.”

Finally, it is quite intelligently argued that there are several constructs of interest that, when measured, should not provide a neutral standpoint on the issue (Presser and Schuman, 1980). For example, when asking a respondent A about their level of satisfaction with service X, they are frequently provided with a bipolar scale, ‘neutral’ being anchored (i.e. the rating-scale has an odd number of intervals), and ‘very satisfied’ anchored at one end with ‘very dissatisfied’ anchored at the opposite end. Presser and Schuman (1980) point out that one who has experienced service X was either satisfied with it or dissatisfied with it. If respondent A was not *satisfied* with service X, then by default, they were *dissatisfied* with it. A ‘true neutral’ does not exist when measuring this construct. Researchers must take note of issues such as these when deciding on whether or not a neutral position should be included.

### 2.3.5 Reliability and Rating-Scale Length

*“The reliability of measurement is a psychometric concept defined as the proportion of the response variance due to true-score variance, where the response variable  $y$  (normally defined as continuous) is defined as the sum of two components: the true value (defined as the expectation of a hypothetical propensity distribution for a fixed person) and a random error score.”* (Alwin, 1992: 89)

Much of the abovementioned literature, in dealing with the number-of-response-categories problem, emphasises *reliability* as the major criterion for recommending the number of response categories to use<sup>1</sup>. However, the debate is further complicated by a disagreement over whether reliability *is* in fact dependant on the number of response categories at all. It would seem that for the most part, researchers believe reliability to be linked with the number of categories adopted (Symonds, 1924, Champney and Marshall, 1939, Jahoda et al., 1951, Ferguson, 1941, Murphy et al., 1938). However, there are a number who argue that reliability is independent of the number of rating-scale points (Jacoby and Mattel, 1971, Bendig, 1954a, Komorita, 1963, Komorita and Graham, 1965, Peabody, 1962).

Bendig (1953a) reported that the reliability of group and individual self-ratings is little affected by variations in the number of rating-scale categories within the limits from 3 to 9, but that both individual and group reliabilities begin to decline when eleven categories are used. He proposed that this drop in reliability beyond eleven categories may have been due to this longer rating-scale presenting the respondent with an

---

<sup>1</sup> Of which either test reliability or respondent reliability, or both, were examined.

introspective problem that is too difficult. Bendig (1953b, 1954a, 1954b) later confirmed the results of his previous study with similar results. In line with Bendig, Komorita (1963) concluded that utilisation of a dichotomous rating-scale would not significantly decrease the reliability of the information obtained when compared to that obtained from a multi-step scale. In addition, Jacoby and Mattel (1971) found that both reliability coefficients (internal consistency and test-retest) were independent of the number of rating-scale intervals used. They further recommended that “reliability should not be a factor in determining a Likert-type scale rating format,” (Jacoby and Mattel, 1971: 498). However, the results from these studies significantly contradict those by others who argue that reliability is an extremely important factor when considering the number of categories to use (Symonds, 1924, Champney and Marshall, 1939).

Some have raised a new perspective to this argument, whereby it has been claimed that respondent-centred variables such as introspective skills, attitudinal set, and degree of involvement of the respondent approaching the task, can have important effects on the reliability and validity of scale scores (Jenkins and Taber, 1977). Jenkins and Taber (1977: 397) criticise that “there is no theory to predict how a respondent’s *behaviour* would change as a function of altering the number of response categories.” Respondent-centred variables and their association with rating-scale length and error variance is an area that requires attention.

### 2.3.6 Validity and Rating-Scale Length

Considering that a scale cannot be valid unless it is reliable, and in view of the fact that some of the abovementioned researchers believe reliability (and in turn the potential for validity) is independent of the number of rating-scale intervals, they are addressed here.

One should question the conclusions by some studies, such as Jacoby and Mattel (1971: 498), who state that “when determining the number of steps in a Likert-scale rating format, validity need not be considered because there is no consistent relationship between it and the number of scale steps utilized.” In light of this, one should consider some contradictory conclusions made by others. For example, Cronbach (1950: 4) first pointed out the effect that response bias<sup>2</sup> has on validity stating that “response sets dilute a test with factors not intended to form part of the test content, and so reduce its logical validity,” before going on to assert that response bias becomes most influential as items become ambiguous or difficult to answer. From this, it could therefore be argued that very fine rating-scale divisions (i.e. longer rating-scales) *do* affect validity by introducing greater error variance in the form of response bias.

Previous research indicates that respondents simplify their rating task by using the range of the response alternatives given to them, as a frame of reference in computing a frequency estimate (Schwarz, 1990, Eiser and Hoepfner, 1991, Tversky and Kahneman, 1974). The underlying assumption behind this estimation procedure is that the

---

<sup>2</sup> Defined as a tendency to respond systematically to questionnaire items on some basis other than what the items were specifically designed to measure (Paulhus, 1991, Greenleaf, 1992a). There are two types or response bias; response set and response style (Cronbach, 1950, Guilford, 1954, Nunnally, 1978). The former occurs when respondents try to present themselves in a certain way, the latter occurs due to the way in which the construct or measurement is presented (Rorer, 1965).

researcher-defined rating-scale (e.g. Likert-type rating-scales or semantic differentials) reflects the researcher's knowledge about the distribution; values in the middle range of the rating-scale reflect the 'average' or 'usual' frequencies, whereas the extreme values of the rating-scale correspond to the extremes of the distribution (Bless et al., 1992). In other studies, respondents report higher behavioural frequencies when the response formats offer high rather than low frequency response alternatives (Schwarz et al., 1985). Related research shows that any variable that may increase task difficulty may also increase respondents' reliance on the range of response alternatives presented to them, resulting in reports that are largely a function of the response alternatives offered (Bless et al., 1992). This is somewhat concerning, as the evidence suggests that the research instrument adopted by the researcher can cloud true variance from respondents by directly affecting the way in which they respond, which in turn affects validity. This reiterates the dilemma of what the 'ideal' rating-scale length is for a given research situation, if it even exists.

Reconciliation of these results is difficult because of the variety of methodological approaches applied across studies. The lack of comparability is particularly acute in that each empirical study uses differing test instruments. Many have concluded that what is most apparent from the extensive body of research is that there is no single number of response alternatives for a rating-scale which is appropriate under *all* circumstances (Cox III, 1980, Jacoby and Mattel, 1971, Garner, 1960). Guilford (1954: 291) put it best by saying, "we are left, therefore without being able to set up any hard and fast rule concerning the number of scale divisions to use." If a hypothetical 'ideal' rating-scale length could be used, then error variance would be less likely to affect the validity and reliability of scores.

Literature links inappropriate rating-scale length<sup>3</sup> with the manifestation of greater error variance, and many argue the largest contributor to this error is response bias (Bardo et al., 1985, Cronbach, 1946, Bardo and Yeager, 1982b, Welkenhuysen-Gybels et al., 2003, Baumgartner and Steenkamp, 2001, Chen et al., 1995). This suggests that response bias seems to be one of the major symptoms of an inappropriate rating-scale length, having a detrimental affect on error variance of scores. Cronbach (1950: 22) stated that the more complex rating-scale offers more chance for personal interpretation, which he linked to response bias tendency, and disagreed with the argument that the longer rating-scale gives more reliability, “since this is precisely what we would expect if all of the added reliable variance were response-set variance and had no relation to beliefs about the attitude-object in question.” On the whole, when rating-scale length is considered with respect to response bias, there have been contradictory conclusions; a greater number of response categories is likely to lead to lower levels of extreme responding (Hui and Triandis, 1989), but more response categories will also lead to higher levels of scale attenuation (Wyer, 1969). These two examples illustrate that the aim of the researcher should be to use enough response categories to avoid encouraging response bias, whilst ensuring that the respondent can clearly distinguish between adjacent response categories in a way that is meaningful<sup>4</sup>.

Whilst numerical anchoring of rating-scales has been examined for its effect on measurement quality, researchers’ choices around verbal anchoring have also been shown to directly affect the quality of measurement.

---

<sup>3</sup> *Inappropriate* in the sense that it has either: too many categories, presenting respondents with an introspective problem; or too few categories, whereby the information transmitting capacity of the rating-scale has not been maximised and the respondent is potentially frustrated.

<sup>4</sup> Such that the respondent finds each response category meaningfully distinct from the adjacent categories.

### 2.3.7 Rating-Scale Verbal Labels

Numbers, words and/or graphic symbols are used to denote the categories on rating-scales, but verbal labelling has become the dominant approach to facilitate communication (Rohrmann, 2003). Either single words or short expressions are used, e.g., "never / seldom / sometimes / often / always", "strongly disagree / disagree / undecided / agree / strongly agree". A typical problem for researchers when choosing appropriate rating-scale labels is how the *amount* and *type* of verbal anchoring of the rating-scale intervals impacts on both the reliability and validity of the scales (Angelmar and Pras, 1978, Bendig, 1953a, Bendig, 1953b, Bendig, 1954a, Bendig, 1954b, Finn, 1972, Wildt and Mazis, 1978, Wallsten et al., 1993, Smith, 2004b). It could be considered that the more defined the rating-scale categories, and the more objective the definitions, the greater the inter-rater measures of reliability will be. However, Bendig (1953a) pointed out that in self-ratings, which are commonly used in personality studies, objective and extensive verbal anchoring may result in an undesirable loss in the 'projective' elements present in such self-ratings. Some reported that rating-scale labels make no difference to the reliability of ratings (Finn, 1972, Peters and McCormick, 1966), stating that "this aspect of rating scale construction would thus appear to be of little consequence," (Finn, 1972: 264). Whereas others found that increased verbal definition of the categories resulted in slightly increased reliability (Bendig, 1953b). What *is* clear, is that *how* rating-scale points are denoted affects response behaviour (French-Lazovik and Gibson, 1984, Hartley et al., 1984, Lehto et al., 2000, Wildt and Mazis, 1978).



Clearly, verbal labelling provides many advantages, such as familiarity, ease-of-explanation, and facilitating the capture of normative judgments. However, this is offset by inferior measurement quality (Rohrman, 2003). The verbal labels selected for inclusion in the rating-scale may not have the same meaning for all respondents, adversely affecting reliability (Jacob, 1971). Additionally, the positions intuitively assigned to each of the terms by the researcher may not correspond to the average perception by the respondents (Angelmar and Pras, 1978).

These problems are compounded when a rating-scale with verbal labels is translated into other languages (van de Vijver and Leung, 1997, Chen et al., 1995, Schaeffer, 1991, Tourangeau and Rasinski, 1988, Angelmar and Pras, 1978). Cross-national comparability of ratings is difficult, as the equivalence of expressions in different languages is usually not known (Angelmar and Pras, 1978, Smith, 2004b). At best, the translated rating-scale will have the same reliability and an identical measurement bias. At worst, reliabilities as well as biases will differ, with obvious effects on comparability (Angelmar and Pras, 1978). This has become an ever-increasing issue in the context of today's globalised world, with more and more research studies crossing national boundaries.

These problems may arise because rating-scales use verbal labels which do not reflect the cognitions of respondents (Rohrman, 2003). Some have set out to avoid this by collecting psychometric and psycholinguistic data in order to create a schema detailing the cognitive position of commonly used labels on a rating-scale continuum for particular populations (Angelmar and Pras, 1978, Rohrman, 2003, Smith, 2004b). This is a way researchers have tried to address the issue of verbal labels being country-

specific. Commonly used adjectives and adverbs are used in tests with respondents to determine the relative magnitude of each word. These studies seek to assess the homogeneity of meaning of these labels, as well as their relative position on cognitive continuums, for a particular population under study. This is done in different ways: terms are ranked in order of weakest to strongest; terms are rated on a numerical scale allowing the distance between terms to be known; and terms are rated on a ratio scale using magnitude measurement techniques. Studies such as these result in having a list of commonly used verbal labels with a guideline as to their relative position on a rating-scale as perceived by a particular culture or population.

Findings from these verbal label studies highlight an important issue. This is best exemplified by one of the results in Rohrmann's (2003) study, where he employed a verbal magnitude scaling task with respondents. On the agreement/disagreement continuum the most commonly used verbal labels (strongly disagree / disagree / neither agree nor disagree / agree / strongly-agree) score 0.4, 1.6, 4.9, 8.2, 9.6 out of ten, respectively, and he points out that they were obviously not fulfilling the equidistant principle<sup>5</sup>. This highlights the problems incurred when verbal labels are used to communicate the middle-categories. It is very difficult for researchers to be able to adequately represent the middle-categories to respondents using appropriate verbal labels. As such, this activity may reduce reliability as opposed to improving it. Bendig (1953a), for example, found that significant improvements to reliability occurred only when the neutral position and end-points were anchored. This would suggest that the inclusion of verbal labels in the mid-range would be overly ambitious and that

---

<sup>5</sup> Rohrmann (2003) argues that if rating-scales are to be constructed which approximate interval scale quality, it is essential to use equidistant scale points. While numbers and/or layout features enhance perceived equidistance, words do not necessarily convey this, and researchers need to address this wherever possible.

researchers should be primarily concerned with verbal labels at the end-points of a rating-scale, and neutral (if included).

The guidelines provided by the abovementioned studies, are intended to be used by researchers to help inform them as to the verbal labels they should use, and their position on the rating-scales when surveying a *particular* population. Whilst this may increase the validity and reliability of the rating-scale when used in that particular research setting, there are still problems. Firstly, there are epistemological issues to be considered. From a cognitive psychology or psycholinguistic perspective, one may question whether a 'universal' (context-free and timeless) meaning of the verbal labels examined at one point in time are valid for the construction of equidistant rating-scales at another point in time, as the way language is used changes over time (Rohrman, 2003). The subjective magnitude and sub-cultural meanings of verbal labels are not constant, even within the same population as was measured the first time. In this context, researchers would have to continuously pretest the population under study for adequacy of rating-scales before every research endeavour. This raises the question of how often this would need to be done, and whether or not this is even practical.

Secondly, it has been argued (e.g. Angelmar and Pras, 1978) that rating-scales be given verbal labels on the basis of empirically devised rating-scale values for the labels in each country under study. However, this overlooks the issue of homogeneity of meaning of the labels *within*-country. If the issue is one of *culture* and its affect on the relative meaning and magnitude of verbal labels, then it is fair to say that countries that have a culturally diverse population (such as the US) may, when surveyed, produce data that contain a greater degree of response error. This complicates the issue further, in that

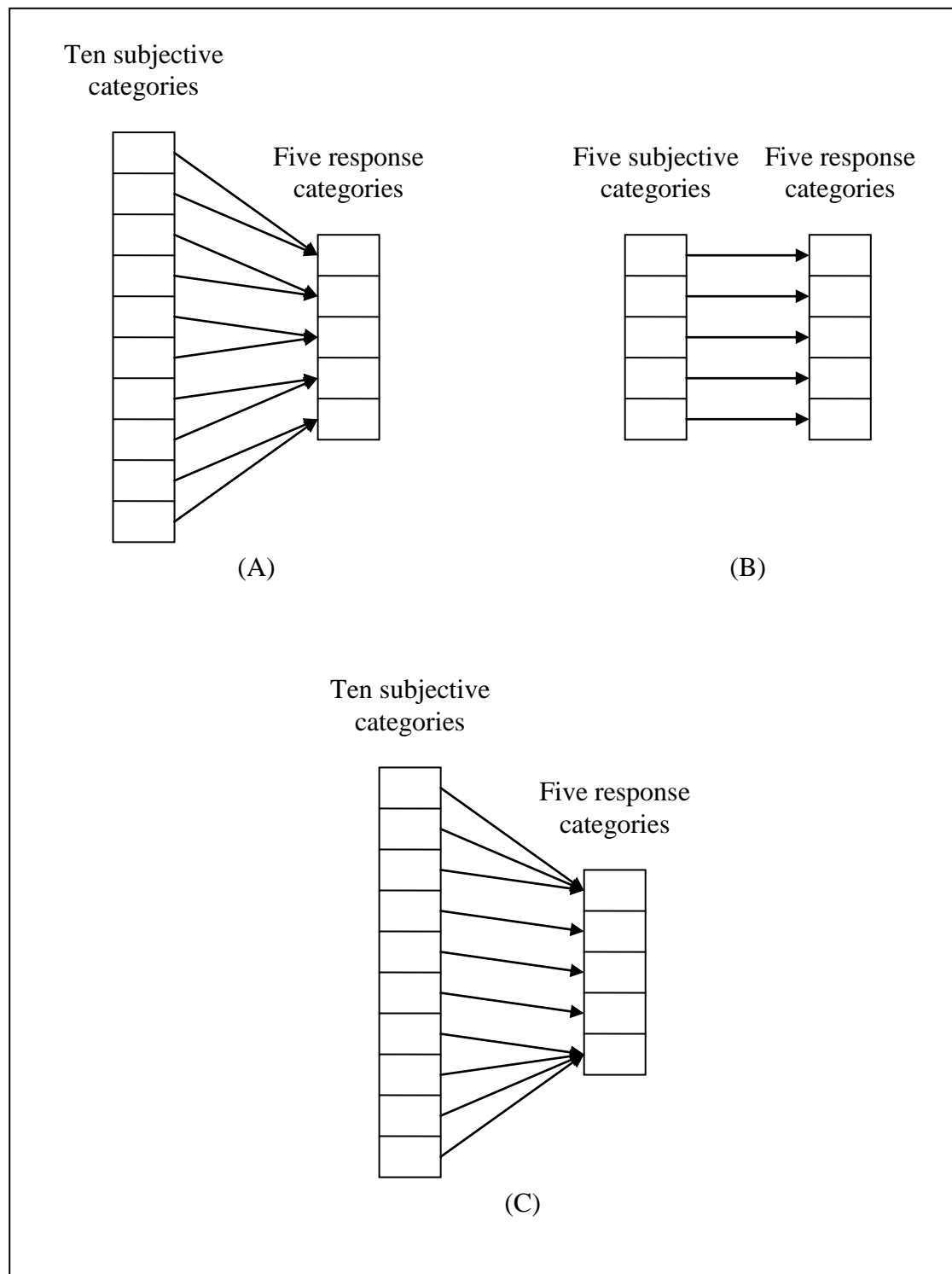
researchers would have to consider the cost to data quality of *not* pretesting to determine appropriate verbal labels and their positions, versus the impracticality of trying to find a rating-scale and verbal labels appropriate to *all* subcultures within a single population through this recommended system of pretesting. The core problem here is that researchers need to be able to create individual-appropriate rating-scales for respondents, and typically this has been attempted through whatever means practical and possible; by grouping respondents together and striving to create culturally-appropriate or country-appropriate rating-scales.

It would be far better for researchers and respondents, if there was some way of ensuring that rating-scales and their verbal labels could remain up-to-date and cognitively accurate (to individual respondents), without a costly and impractical means of achieving this (for researchers). Data quality would be enhanced, and so too would the experience of all parties involved.

## **2.4 Response Bias**

It has long been established that inappropriate rating-scale length and format has an effect on the quality of data collected (Bendig, 1954a, Bendig, 1954b, Benson, 1971, Finn, 1972, Green and Rao, 1971, Komorita and Graham, 1965, Jacoby and Mattel, 1971). Of particular interest, given its gross impact on data quality, is the fact that rating-scale length has been linked to the manifestation of *response bias* (Javeline, 1999, Hui and Triandis, 1989, Bardo et al., 1985). This link was uncovered by studies on human judgment and how judgment is mapped onto a rating-scale (Wyer and Carlston, 1979). A distinction is drawn between the subjective categories of judgment and response categories (or rating-scales). The former are present within a respondent's

mind and are used to process incoming information; “I love it”, “That’s OK” are examples. The latter are those provided by the researcher, and are therefore researcher-defined rating-scales (or response categories). “When subjects respond to a rating scale, they begin by mapping their somewhat “elastic” individually held subjective categories onto the response categories” (Hui and Triandis, 1989: 298). The latter have anchors that may approximate the meanings of some of the subjective categories, see (A) and (B) in Figure 2. 1.



**Figure 2. 1 Mapping of subjective categories on response categories.**  
**Source (Hui and Triandis, 1989).**

As such, the appropriate number of subjective categories depends on the individual respondent (for greater clarity, the word 'subjective' will be replaced with the word 'ideal' hereafter). This emphasises that what is 'ideal', in terms of number of categories,

varies by individual. (A) in the previous figure shows optimal mapping when number of ideal categories exceeds number of response categories. (B) shows the mapping when number of ideal categories equals number of response categories. If respondents fail to contract or stretch their ideal categories adequately to match the response categories, they will respond differently than someone possessing the same attribute who easily manipulates their ideal categories (Hui and Triandis, 1989). (C) in Figure 2. 1 provides an illustration of suboptimal mapping when the number of ideal categories exceeds number of response categories. Here, Hui and Triandis (1989) argued that respondents who engage in extreme responding (one of several types of response bias), are those who have more ideal categories that represent great intensity (in either direction) than is allowed on the rating-scale. Thus, respondents solve this problem by mapping several ideal categories onto the same extreme response category, resulting in frequent checking on the endpoints of the rating-scale. In this case, the addition of more categories to the rating-scale would potentially simplify the task for respondents who cannot “elastically” and more evenly distribute their ideal categories over the response categories. This example is a good illustration of how inappropriate rating-scale length can provoke response bias contamination in survey data.

This issue has also been referred to as a question of *meaningfulness* of response categories and their relationship to the manifestation of response styles (Viswanathan et al., 2004). In this context, Viswanathan et al. (2004: 109) defined meaningfulness as “the number of categories that individuals typically use in thinking about an attribute in such situations as making a choice or judgement.” In this way, the response categories are likely to be more *meaningful* to respondents if they closely match the respondents’ ideal categories of judgment.

It is recognised by many that response bias can influence measures of abilities, attitudes, opinions, beliefs and personality (Couch and Keniston, 1960, Cronbach, 1946, Cronbach, 1950, Hamilton, 1968, Oskamp, 1977). Oskamp defines response sets as “systematic ways of answering which are not directly related to the question content, but which represent typical behavioural characteristics of the respondents” (1977: 37). Couch and Keniston (1960: 151) make two points when referring to response styles: the first is that response styles are “primarily a statistical nuisance that must be controlled or suppressed by appropriate mathematical techniques,” the second treats response styles as “a manifestation of deep-seated personality syndrome[s].”

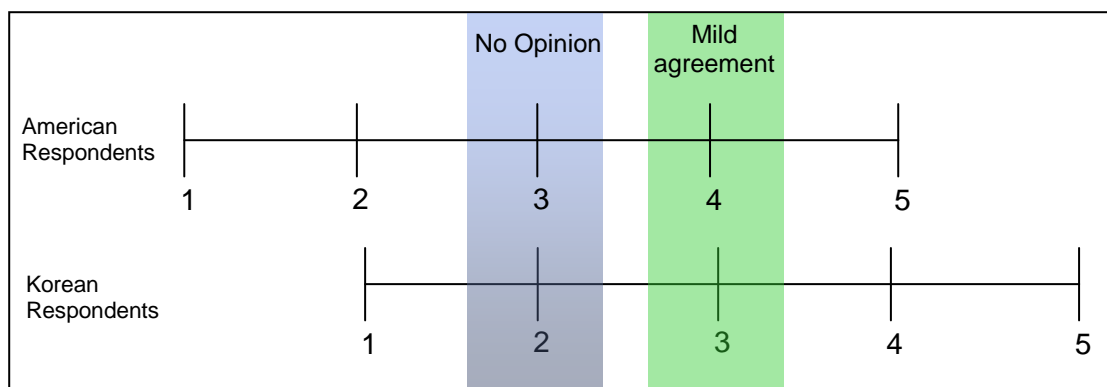
#### **2.4.1 Response Bias and Data Analysis**

It is important to emphasise the impact that response bias has on data analysis in order to appreciate why having an appropriate rating-scale is so important if data quality issues are to be minimised. Many different response biases have been identified. Diamantopoulos et al. (2006) provide a good summary for how response styles may be manifested in data, describing that they relate to:

- The respondent’s tendency to use *particular points* on a scale, such as extreme responding, mid-point responding and the tendency to rate to the left (or right) of centre (Presser and Schuman, 1980, Greenleaf, 1992b);
- The respondent’s *spread of responses* (response range and index of dispersion) such as scale attenuation (Wyer, 1969, Greenleaf, 1992a);
- The respondent’s *reaction to direction of the item or response category* such as (dis)acquiescence response style (Couch and Keniston, 1961, Bachman and O’Malley, 1984).



Looking at how rating-scale length affects response bias, with some groups too few response categories manifest as extreme responding (Hui and Triandis, 1989), too many categories result in scale attenuation which results in low index of dispersion scores (Wyer, 1969). Acquiescence/disacquiescence could be a result of respondents inadequately mapping their ideal categories of judgement onto the response categories, thereby causing an imbalance in the operative use of the rating-scale. Whilst this has not been empirically tested, it seems theoretically plausible given what is already known. For example, Riordan and Vandenberg (1994) found that a response of 3 on a 5-point Likert-type rating-scale (LTRS) means “no opinion” to American respondents but “mild agreement” to Korean respondents. This effectively causes a misalignment in the use of the rating-scale across groups, illustrated in Figure 2. 2.



**Figure 2. 2 Adapted from the findings of American versus Korean’s use of rating-scales in a study by Riordan and Vandenberg (1994)**

From this example, the Koreans may appear to be suffering from a greater tendency to disacquiesce, if researchers are working on the assumption that the reaction to the construct under measurement is stable across both groups and that the distribution of respondents’ position on the construct is equivalent between groups. This would be because on the whole, when the Koreans wanted to report *no opinion* on an issue, they would be perceived by the researcher to be *disagreeing* with it. In order for it to appear

as though they were using the rating-scale relatively uniformly, their true opinion would have to *agree* with the items to a greater extent than did the Americans. This particular example illustrates the point about rating-scale length and its likely affect on acquiescence/ disacquiescence, but it also happens to highlight the issue of cross-cultural differences in rating-scale use. Whilst the issue of culture will be further explored in a later section, this example was highlighted here as it emphasises that cross-cultural studies have identified issues relating to response behaviour and rating-scales that are less obvious within mono-cultural environments. These issues are likely to exist as *individual-level* differences within mono-cultural environments, but so far research has only highlighted these differences at the national/cultural/sub-cultural level.

Researchers should be very concerned that inadequate rating-scales have been linked to response bias, given that response bias greatly reduces the validity of research findings (Broughton and Wasel, 1990). Specifically response bias;

- Can contaminate observed responses by either inflating or deflating respondents' scores on measurement instruments (Bagozzi, 1994).
- Can affect conclusions about the relationship between scales by either inflating or deflating the correlation between respondents' scores on measurement instruments (Bagozzi, 1994).
- Can make it appear that there are differences between groups when no differences actually exist and/or can hide real differences between groups (Heide and Gronhaug, 1992, Baumgartner and Steenkamp, 2001).
- Can increase the association between variables to the extent that significant relationships appear, yet can also decrease associations to the extent that a

relationship is not revealed (Chun et al., 1974, Lorr and Wunderlich, 1980, Bardo and Yeager, 1982b, Heide and Gronhaug, 1992).

These make a particularly strong argument for the creation of a rating-scale that is cognitively appropriate to the ideal categories of respondents. Given response bias includes unconscious tendencies within the respondent (Baumgartner and Steenkamp, 2001) to react to situational effects of the survey (e.g. response format, survey length, type of items etc) in a particular way, a problem resulting from an inability to map one's ideal categories onto the response categories would further exaggerate the response bias problem (Hui and Triandis, 1989). If the negative consequences of a mapping mismatch, in particular response bias, could be avoided, then researchers would have more confidence in the quality of their data and the conclusions drawn from their research.

#### **2.4.2 Existing techniques to measure and correct for response bias**

Whilst it is true that researchers and methodologists have been developing techniques to measure and correct for response bias in survey data, they are few and prone to criticism (Cornwell and Dunlop, 1994, Closs, 1996, Van Hemert et al., 2002, Smith, 2004a). Researchers generally use one of a small range of methods to measure the amount of response bias present in data. One method, is the inclusion of uncorrelated items to examine the spread of responses to those items, whereby the spread would be expected to be uniform if no response bias is present (Greenleaf, 1992b). The problem with this method is that it assumes that uncorrelated items in one group or context will remain uncorrelated in another. The researcher cannot always be completely confident in the uncorrelated nature of the items. Additionally, survey length is increased by the

inclusion of additional items with the associated increase in costs and decrease in response rates (Smith et al., 2003), and this might introduce other biasing effects.

Another approach is to collect both attitudinal and behavioural information as response bias has less of an impact on more concrete (i.e., behavioural) information (Greenleaf, 1992a). While this method is possible across groups, it is not ideal as (a) it can be difficult to develop behavioural measures that are directly related to attitudinal constructs (look, for instance, at the relationship between attitudinal loyalty and repurchase behaviour (Fazio et al., 1989)); and (b) the need to collect two types of information from all respondents again increases the length of the survey.

A further method is to estimate the impact of response bias on one measurement scale from the way in which the subject responds to the items on all of the other measurement scales on a research instrument (Baumgartner and Steenkamp, 2001). This method is based on the premise that while respondents might truly have extreme views in some areas, they are unlikely to have extreme views in all areas (Couch and Keniston, 1960). Yet if a respondent truly has an extreme opinion on one construct, it is not unreasonable to believe that they will also have extreme opinions on related constructs (Couch and Keniston, 1960). As such, to use this method to estimate the impact of response style bias, the assumption needs to be made that the constructs being measured are independent. This assumption is likely to be violated in both academic and commercial research.

Finally, some have argued for the use of Item Response Theory, such as de Jong et al.'s (de Jong et al., 2008) IRT-based model, to measure response style. However, IRT

models in cross-group research require measurement-invariant anchor items to make the scale of the latent variable common across groups (May, 2006). The item parameters need to be the same across groups for these anchor items in order for the model to be identified. Conceptually, this is likely to be violated in settings such as cross-cultural research. De Jong et al. (2008), with their IRT-based model, may have overcome this as well as other issues with previous IRT approaches. However, there is still a method that (a) focuses predominantly on ERS, meaning that the data would still potentially be contaminated by other types of response style, and (b) requires researchers to have an advanced statistical knowledge in order to understand and apply a complex and time-consuming method of data correction.

Other forms of post-hoc statistical techniques have been used to remove the impact of response bias. For instance, the use of ANCOVA and partial correlations (Diamantopoulos et al., 2006), ipsative rescaling (Cunningham et al., 1977, Gurwitz, 1987, Broughton and Wasel, 1990), and standardisation (Hofstede, 1980, Fischer, 2004). However, many of these techniques have attracted criticism as a result of their impact on survey data. For example, reliabilities have been shown to be seriously inflated (Tenopyr, 1988) and deflated (Bartram, 1996). Data that has been transformed through standardisation have been said to be participant specific, which makes participant responses incommensurable (Horton, 1974, Stewart, 1981). When using ipsative rescaling, the meaning of the original responses have been said to be fundamentally altered (Gurwitz, 1987, Cheung and Rensvold, 2000). This can make interpretations difficult and render the results of analyses invalid, for example in factor analysis spurious correlations are imposed among the items (Closs, 1996, Baron, 1996).

Generally, therefore, it would seem that that transformed data have to be interpreted with great caution.

Despite citing eliminating potential bias as a reason for employing some of the statistical correction techniques raised above, most researchers do not explicitly discuss why observed mean differences constitute bias rather than substantial variation that might be linked to their topic of interest. For example, Smith (2004a) and Van Hemert et al. (2002) pointed out that differences in response bias can be explained in terms of psychological dimensions, such as the ones derived from Hofstede's (1980) work. Therefore, these differences might better reflect different communication styles rather than bias that needs to be controlled for. In fact, some researchers have shown that response patterns may even be a form of communication style related to cultural (or personal) characteristics (Van Hemert et al., 2002, Smith, 2004a). According to this view, some statistical techniques (such as standardisation) remove variation that is substantial and relates to individual communication.

The fundamental flaw of these statistical correction techniques is that the corrections are post hoc. They require the researcher to determine the amount of response bias present in the observed score without being aware of what the true score actually is. Given that the measurement and correction of response bias is clearly problematic, it would therefore seem reasonable to try to 'design out' the possibility of response bias *before* the data collection process. In this way, situational factors (specifically rating-scale length) can either discourage or encourage a person's inherent tendency to engage in stylistic responding (Snyder and Ickes, 1985, Baumgartner and Steenkamp, 2001). This would mean that the measurement choices taken by the researcher would directly

impact on the data collected from respondents. In the past it was not considered possible to completely eliminate response bias through survey design (Couch and Keniston, 1960). In light of a changed external environment, it is now theoretically possible to examine the problem at an individual-level; technology can be used to re-examine the design problem. Therefore, it may now be possible to reduce such problems caused by inappropriate rating-scale length, and thus minimise its impact on substantive conclusions.

## **2.5 Cross-Cultural Measurement Equivalence**

As was briefly mentioned earlier, culture has been shown to relate to response bias. It plays a role in the way respondents communicate their responses through measurement instruments, with some researchers concluding that respondents from certain cultures are more prone to adopting certain response styles (Gibbons et al., 1999, Hui and Triandis, 1989, Ellis and Kimmel, 1992, Javeline, 1999, Si and Cullen, 1998). However, it is potentially misguided to conclude that a respondent's culture (as an independent variable) has implications on the type or degree of response bias manifested (as the dependant variable). For it could simply be that the measurement instrument (independent variable), and specifically the rating-scale length, was inappropriate to the respondent from that particular culture, thus catalysing the manifestation of response bias (dependant variable). This is analogous to one trying to measure how much someone weighs in kilograms, and they assume that you are asking for their weight in pounds. This raises the issue of rating-scales and their relationship with *measurement equivalence*.

One of the biggest issues with multinational research is whether similarities or differences are in fact *real* (Barksdale and McTier-Anderson, 1982, Harkness et al., 2003a). Some researchers have questioned whether measurement problems inherent in international research have attenuated results that are different from what was expected, that is, whether the results are measurement and scaling artefacts or true cultural differences (Adler et al., 1989). In order to reduce threats to reliability and validity, for some time researchers have been instructed to address the problems of measurement equivalence (Adler et al., 1989, Albaum and Peterson, 1984, Davis et al., 1981, Aulakh and Kotabe, 1993). Research recognises that scale designs suitable for one population may not be suitable for other cultural groups (Harkness, 2003c). In part, researchers seek to address this problem through the development of statistical techniques that cater to equivalence issues in survey data. Whilst useful to researchers, it can be argued that they suffer from the same drawback of being applied post hoc as those statistical techniques already mentioned. Van de Vijver (2003: 233) sums up the issue quite nicely:

*“The statistical toolbox of the cross-cultural survey researcher has become both larger and more sophisticated in the last few decades, [...] It is important to point out that comparative techniques for instrument design have not enjoyed anything like the same degree of refinement. It would be fatal to neglect design and to expect analysis to produce a silk purse and the wherewithal to fill it. Statistical sophistication in data analysis cannot compensate for poor quality of study design nor for lack of cultural sophistication. [...] Both a sophisticated analysis of a poor instrument and a poor analysis of a good instrument yield low quality.”*



Measurement equivalence addresses the question of whether the same models hold across different populations, in other words, whether the measures are comparable (Harkness et al., 2003a). Douglas and Craig (1983) divide measurement equivalence into three overlapping areas: *calibration*, *translation* and *metric equivalence*. Both calibration and translation equivalence seek to ensure that measurement instruments mean the same thing after translation, with the former being concerned with whether the units of measurement are the same in different countries, and the latter implying that the same questionnaire items measure the same latent constructs in different populations (Mullen, 1995). However, the underlying assumption of this is the behaviourist stimulus-response approach to survey design.

*“Though establishing equivalency in question concept and substance is very important, it is only half the battle. Surveys also need equivalency in response categories.”* (Smith, 2003b: 73-74)

The traditional stimulus-response approach to survey design presupposes that rating-scale verbal labels can be directly translated into another language and be perceived to be equivalent (see earlier discussion on verbal labels, Section 2.3.7). The main focus for translation equivalence in this context has been on the translation of scale *items* and not rating-scale *labels*. In practice, this is probably because rating-scales for source questionnaires are rarely specifically 'designed' for comparative projects in other populations (Harkness, 2003c). Source questionnaire designers use the rating-scales with which they are familiar, either from their own languages and survey traditions or from the language in which the questionnaire is being developed. These are automatically culturally and linguistically anchored in the source language (Harkness, 2003c). Given the shift away from looking at the problem through a stimulus-response lens, these issues should no longer be overlooked.

In order for *metric equivalence* to exist, the psychometric properties of data from multiple groups must exhibit the same coherence or structure (Berry, 1980). In other words, subjects must respond to the rating-scales in the same way. Douglas and Craig (1983) highlight two threats to metric equivalence: inconsistent scoring across populations and scalar inequivalence. Inconsistent scoring poses a threat to the reliability of measurements (Davis et al., 1981, Douglas and Craig, 1983, Bhalla and Lin, 1987). Inappropriate rating-scale length has already been shown to affect the reliability of measures (see discussion in Section 2.3.5), and therefore result in inconsistent scoring. Scalar equivalence and response bias also threaten metric equivalence (Cunningham et al., 1977, England and Harpaz, 1983), in that the scores obtained from respondents in different countries may not have the same meaning and interpretation (Douglas and Craig, 1983). Whilst these scores may differ due to cultural characteristics (Vijier and Poortinga, 1982, Chen et al., 1995, de Jong et al., 2008), a confounding factor may have been that of a grossly inappropriate rating-scale length for the group in question. These differences in scalar equivalence/response bias result in the addition of error to measurements, threatening the validity of cross-national comparisons (Mullen, 1995).

However, there are other hazards that are derived from differences in respondents' information processing, that is, the particular way in which a group cognitively operates (Cheung and Rensvold, 2000, Little, 2000). Styles of decoding and encoding information when answering items differ across some cultures (Arce-Ferrer and Ketterer, 2003). Cross-cultural research calibration is an issue *between* groups. However, there appears to be an assumption with mono-cultural research that

calibration equivalence is present between individuals and small subgroups. Whilst it is a lot easier for researchers to observe differences in response behaviour *between* cultures, it is harder to quantifiably notice *individual-level* differences *within* cultures. It is said that,

*“issues that might be able to be ignored in monocultural contexts cannot, however, be ignored in cross-cultural research. Comparative researchers have no grounds to assume identity of meaning across social, linguistic, or cultural groups.”* (Harkness et al., 2003a: 8)

However, it can also be argued that those issues are wrongfully being ignored in a mono-cultural context, especially if they are impacting the substantive conclusions drawn from the data. At an individual-level, researchers assume identity of meaning, and this may also introduce serious bias. If this individual-level variance in response behaviour is substantial enough, the issue no longer becomes one of cross-cultural/cross-national measurement equivalence but one of cross-respondent measurement equivalence. The development of sound (well-developed and well-tested) new instruments has been called for to reduce problems of equivalence in the cross-cultural context (Harkness et al., 2003b: 29). However, these measurement issues are also highlighted within *mono-cultural research*. Here researchers are turning to tailoring aspects of the survey process with a view to enhancing response, and thus data, quality (Harkness et al., 2003a: 9).

This is very indicative of the shift from the stimulus-response approach to survey design to a more cognitive science. It is a growing school of thought, with one of its interests examining how individual-level differences impact the quality of data (Sirken et al., 1999, Rossiter, 2002). It is through this new lens that the issue of calibration should be

considered. Whilst individual-level differences are exaggerated when comparisons are made across cultures, this does not mean to say that substantial differences do not exist *between* individuals from within a culture.

## **2.6 Individual Characteristics**

Theoretical explanations of response behaviour are of either the dispositional or situational variety (Baumgartner and Steenkamp, 2001, Snyder and Ickes, 1985). In the literature, rating-scale length, a key situational factor, is examined for its affect on response behaviour. There are also studies that have identified relationships between response bias and dispositional factors (individual characteristics such as personality (Cronbach, 1946, Cronbach, 1950, Crandall, 1982, Berg and Collier, 1953, Iwawaki and Zax, 1969, Lewis and Taylor, 1955, Merrens, 1971, Norman, 1969, Zax et al., 1964); age (Osgood et al., 1957); education (Light et al., 1965); gender (Berg and Collier, 1953, Lewis and Taylor, 1955); culture (Smith, 2004a, de Jong et al., 2008); and occupation and social class (see Hamilton, 1968 for a summary of these studies)). The findings are varied and some appear contradictory. However, given that there appears to be a causal relationship between rating-scale length and the manifestation of response bias, and that there is an empirically proven relationship between response bias and personal characteristics, it would seem extremely plausible that there exists a relationship between certain personal characteristics and ideal rating-scale length (ideal number of response categories). As such, whilst rating-scale length preferences and their relationship with individual characteristics have not been investigated extensively, previous studies on response bias and individual characteristics can provide clues as to the likely associations. Where response bias occurs, it is assumed that rating-scale

length/format is not 'ideal', and as such this may provide some insight as to the types of respondent characteristics that may relate to rating-scale preferences.

In the past, various researchers have supported the notion that the developmental level of an individual would play a role in the manifestation of extreme responses (Werner, 1948, Lewin, 1951, Zax et al., 1964). These writers share the idea of progression from a less to a more differentiated psychic structure. They argue that as categories of experience multiply, there is more "freedom of choice" for making judgments. From this, a young child, would be more likely to give all or nothing reactions than would an older child and this has been observed (Light et al., 1965). So, it could be argued that children have very few ideal categories when making judgments. Indeed, Werner (1948) viewed this difference between adult and childhood intellect as being due to the gradual development of the 'abstract faculty'. Light et al. (1965) found as a result of their study, that differences in extreme responding were found as a function of IQ (the lower the IQ, the greater tendency for extreme responding), and stated that this was not surprising as IQ is often related to one's ability to use abstractions. It would seem plausible that those with a lower IQ had fewer ideal categories of judgment, and would therefore react to response categories (with too many intervals) by bunching their ratings at the extreme endpoints (scale attenuation), given their inability to attach meaning to the middle-categories. This would echo Hui and Triandis' interpretation (1989) of what occurs during a mis-mapping of one's ideal categories to the response categories provided.

Gender and the manifestation of response bias have been examined in previous studies with contradictory results. For example Light et al. (1965) found no differences in the tendency for extreme responses between males and females, which contradicts some

previous studies (Berg and Collier, 1953, Osgood et al., 1957, Hamilton, 1968, Bresnahan et al., 1999, Crandall, 1973). However, this may be as a result of their sample comprising only children, whereas studies that *have* shown gender differences in the manifestation of response bias have had samples with adult respondents. This could imply that there may be a difference between males and females and their ideal rating-scales.

All findings considered, the studies in question did not use exactly the same methodologies and so different conclusions are inevitable to some degree. What *can* be said is that there appears to be a link between individual characteristics and the manifestation of response bias, implying there is also a potential link between these individual characteristics and ideal rating-scales.

### **2.6.1 Cognitive approach to survey methodology**

Research recognises that rating-scale designs suitable for one population may not be suitable for other groups (Harkness, 2003c). When researchers choose a standardised rating-scale a priori for use in a survey, their decision is sometimes informed by what they know and can observe about the target population. These observables include respondent demographics such as those previously discussed; age, gender, level of education, culture. This observable information helps researchers to select particular rating-scales for inclusion in a survey. For example, where the sample population is likely to have low levels of education and be younger, conscientious researchers would probably choose to opt for rating-scales with fewer response categories. We know that previous approaches to determining standardised rating-scales were based in the behaviourist philosophy of survey research (i.e. stimulus-response) (Schwarz, 1999).

However, survey researchers have long recognised that individual differences impact on survey administration and this needs to be taken into account when considering rating-scale choices in survey research today (Hui and Triandis, 1989, Ory and Poggio, 1981).

Theoretically, if researchers did not have to standardise a rating-scale across individuals, and it was possible for respondents to use their ideal rating-scales when rating, relationships between individual traits and rating-scale length, seem likely. Given individual traits have been linked with response bias (Cronbach, 1946, Cronbach, 1950, Crandall, 1982, Berg and Collier, 1953, Iwawaki and Zax, 1969, Lewis and Taylor, 1955, Merrens, 1971, Norman, 1969, Zax et al., 1964), and response bias is a manifestation of an inappropriate rating-scale (Bardo et al., 1985, Hui and Triandis, 1989, Javeline, 1999), it is a natural proposition that there be a relationship between a respondent's traits and their ideal rating-scale. It might therefore be of interest to examine particular traits in terms of their relationship to response behaviour. Whilst these traits would be unobservable to researchers prior to a survey, and therefore would not be useful in instructing researchers how to better select a standardised rating-scale for a given population, they would help us to understand more about what goes on in the mind of a respondent. Should relationships between individual traits and ideal rating-scale length be established, then this would contribute to the ever increasing body of research that examines the cognitive aspects of survey methodology (CASM) (O'Muircheartaigh, 1999, Sirken et al., 1999).

When researchers take a behaviourist approach to survey methodology, they are not interested in the underlying processes between stimulus and response.

*“In general, models of response errors in surveys focus on the task, which is constrained and structured to accomplish the research goals – in particular to provide the data necessary for analysis. [...] The respondent is largely disregarded, seen as an obstacle to be overcome rather than an active participant in the process.”* (O’Muircheartaigh, 1999: 43)

The cognitive approach to survey methodology indicates that there are many intervening steps between creating a stimulus and recording a response which can be influenced by individual characteristics. The shift in survey response research from the behaviourist to the cognitive paradigm implies that the two-stage stimulus-response sequence is intersected by a cognitive phase in which respondents perform a series of mental tasks in responding to survey questions (producing a three-stage stimulus-cognition-response model). To arrive at a meaningful response, survey respondents need to perform a series of tasks (Schwarz et al., 1985, Tourangeau, 1984, Strack and Martin, 1987), outlined in Table 2. 1.



**Table 2. 1 Survey response process based on the cognitive paradigm, (for greater detail refer to Sirken et al., 1999).**

Stage in response process	Activities involved
Question is administered	<ul style="list-style-type: none"> <li>▪ Respondent perceives the question.</li> </ul>
a) Interpret the question	<ul style="list-style-type: none"> <li>▪ Understand what is meant.</li> <li>▪ Determine which information to provide.</li> </ul>
b) Generate an opinion	<ul style="list-style-type: none"> <li>▪ If it is an attitude question, either retrieve a previously formed attitude or compute a judgment on the spot.</li> <li>▪ If it is a behavioural question, retrieve instances from memory.</li> <li>▪ Private judgment is formed in the mind.</li> </ul>
c) Respond to question	<ul style="list-style-type: none"> <li>▪ Potentially have to format judgment to fit the response alternatives provided.</li> <li>▪ May wish to edit response before communicating it.</li> </ul>

Researchers that adopt a cognitive approach are more interested in how ‘a’ leads to ‘b’ and leads to ‘c’, rather than simply the outcome. The assumptions in Table 2. 1, from cognitive science, have the potential of theoretically fortifying the practice of questionnaire design in addition to providing practical tools (Graesser et al., 1999). Schwarz (1999) points out that although it is conceptually useful to present respondents’ tasks in this manner, respondents’ actual performance may deviate from this ordering and they may, for example, change their interpretation of the question once they find it difficult to map their answer onto the response alternatives provided by the researcher (Schwarz et al., 1985). Context effects such as these can be classified according to the component of the response process affected by the question context: the impact of context on the interpretation of the target question; the information retrieved in answering it; the use of that information in judging the target issue; or the reporting of the judgement (Tourangeau, 1999). Of particular interest is how context can alter how

respondents map their judgments onto the rating-scale and how they edit their answers before reporting them. Individually *inappropriate* rating-scales adversely affect this process, and it is here that the notion of respondents being able to use their own ideal rating-scales is posited.

Empirical studies conclude that various facets of personality and individual traits are linked with response bias (Iwawaki and Zax, 1969, Couch and Keniston, 1960, Hamilton, 1968). However, other studies found that those same traits are *not* related to response bias (Grimm and Church, 1999) with one author stating that “personality variables previously thought to be related to extreme response style may have been specific to the assessment methods used” (Merrens, 1971). This further emphasises the issue as one of appropriate rating-scales. Whatever the relationship between individual traits and response bias, it is of greater use to consider whether there is a relationship between individual traits and respondents’ ideal rating-scales. There has not been any prior significant research in this area, thus there is no directly-related literature to draw upon for the generation of hypotheses. Given that the evidence from the response bias literature is contradictory in terms of the relationships with individual traits, it may not be fruitful to draw heavily from that literature when conceptualising the factors that could impact on respondents’ ideal rating-scales. As such, it may be better to take a more exploratory approach and start from broader based individual traits such as: how individuals *think*; how *emotion* impacts on their judgment; their need for *structure*; and how people are categorised on broad-based personality traits.

## 2.7 Individualised Rating-Scales

*“A response scale should fulfil psychometric standards of measurement quality as well as practicality criteria, such as comprehensibility for respondents and ease of use. Rating scales are so popular because of their convenience [...] but they are also questionable because of serious shortcomings in their measurement features.”* (Rohrmann, 2003: 2)

Response formats need to be further developed so as to overcome some of the inherent measurement problems. Innovations can be best achieved if we consider the role of the respondent as an active part in the data capture process, as per the cognitive approach to survey methodology. In fact, scale development theory has already broadened Churchill’s (1979) scale development paradigm to include the role of the respondent. Rossiter’s (2002: 319) C-OAR-SE scale development procedure, for example, highlights the need to consider how the respondent affects the construct under study; “the rater entity is an intrinsic component of a marketing construct [...] and largely determines how reliability (precision of scale scores) should be assessed and reported.” This way of looking at the issue is grounded in a shift from the behaviourist approach to survey methodology to a cognitive one, fuelled by movements such as CASM (Conrad, 1999, Graesser et al., 1999, Groves, 1999, Herrmann, 1999, O’Muircheartaigh, 1999, Schober, 1999, Schwarz, 1999, Sirken and Schechter, 1999, Tourangeau, 1999).

There are several researchers that have argued the benefits of involving the respondent in the generation of more *meaningful*<sup>6</sup> rating-scales, with some investigating respondents’ ability to self-anchor a rating-scale (Nugent, 2004, Kilpatrick and Cantril,

---

<sup>6</sup> One can view *meaningful*, in this context, as the degree of similarity between one’s ideal categories of judgment and the response categories provided.

1960, Battle et al., 1966, Donnelly and Carswell, 2002). Theoretically, it can be argued that there are three key ways in which a respondent can self-anchor a rating-scale:

- *Verbally* anchor the scale endpoints, by attaching verbal labels to qualify intervals or endpoints.
- *Numerically* anchor the rating-scale, by attaching numerical values to endpoints (i.e. defining the number of response categories they would like to use),
- *Conceptually* anchor the scale endpoints, by conceptualising intervals or endpoints through association with a particular stance/state.

In theory, these methods can be used in conjunction with one another or independently. For example, a respondent can be asked to *verbally* anchor the endpoints to a rating-scale that already has fixed numbered intervals (i.e. numerical endpoints are researcher-defined). Or, for example, a respondent can be provided with two verbal endpoints and asked to anchor the *numerical* endpoints that correspond to those fixed verbal anchors. Previous research has experimented with allowing respondents to personalise *conceptually* anchored rating-scale endpoints, in that the numerical endpoints are shown to the respondent (and are fixed), and they are then asked to anchor the two extreme endpoints with a meaningful scenario specified by the researcher. For example, Nugent (2004), when looking at social work practice, involved respondents in the design of the rating-scales used to measure their level of depression. Specifically, Nugent (2004: 171) asked

*“the client to imagine a thermometer-type instrument that measures the magnitude of [...] depression, with higher scores indicating a greater intensity problem with depression and lower scores indicative of a lower magnitude problem.”*

Respondents imagined what, for them, were their maximum and minimum depression intensities and then indicated their current state of depression using that rating-scale. Bloom et al. (1999) defined this type of self-defined rating-scale as an ‘individualised rating-scale’. These types of individualised rating-scales have been shown to be reliable (Battle et al., 1966, Morrison et al., 1978) and valid (Battle et al., 1966, Bond et al., 1979, Mintz et al., 1979). Thyer et al. (1984), for example, combined the measurement of the subject’s individualised rating, the researcher’s systematic recording of overt behaviour, and the objective assessment of relevant physiological variables (e.g. heart rate) and found significant correlations between the subject’s own rating and the two physiological indices. Similarly, Nugent’s (2004: 117) results “suggested that a two-item self-anchored scale could serve as a valid unidimensional measure of the construct of depression.”

Kilpatrick and Cantril (1960) describe their self-anchoring scale approach as one in which each respondent is asked to describe, in terms of his own perceptions, the top and bottom of the dimension on which scale measurement is desired, and then to employ this self-defined continuum as a measuring device. They argued that,

*“Since each of us behaves in terms of his “reality world,” the only world he knows, it follows that the key to an understanding of human behaviour is to take into account the unique reality world of the individual. This we have characterized as adopting the first-person point of view, as opposed*

*to the third-person point of view which assumes an objectively definable reality which, except for error is the same for all.*" (Kilpatrick and Cantril, 1960: 1)

For these authors, self-anchored scaling was an attempt to apply this first-person approach to the measurement of psychological variables. They argued against the employment of rigidly predefined dimensions, verbal categories, prepared phrases or sentences, adjective check lists and the like. The unique perceptions, goals and values of each individual were taken into account through the measurement method itself. For example, respondents were asked to describe what, for them, would be the very best or ideal way of life and also what, for them, would they perceive to be the very worst way of life for themselves. Their responses to both were recorded verbatim. The respondent was then handed a pictorial, non-verbal scale, such as a ladder with ten rungs, and told that the top and bottom of the ladder represent the two previously anchored concepts. In reference to verbal labelling, Kilpatrick and Cantril (1960: 3) argued that,

*"The whole point of the method is that the scale is a self-defined continuum anchored at either end in terms of personal perception. The introduction of verbal tags along the continuum would destroy this concept".*

In this case, the authors conceptually anchored the rating-scale at fixed states. But they prompted respondents to personalise these anchored concepts, rendering the endpoints more meaningful on a rating-scale that has a fixed number of intervals (albeit in the form of the rungs on a ladder) and no verbal anchors (fixed or otherwise).

These studies demonstrate that respondents can take a rating-scale that has been *conceptually* anchored by the researcher and render it more meaningful when prompted

to do so. Add to this a scenario where respondents *verbally*-anchor and *numerically*-anchor their own rating-scales; this could further maximise the meaningfulness of the rating-scale. These different ways of anchoring have previously been used in isolation (Finn, 1972). However, there are no studies that demonstrate the use of both respondent verbal and numerical anchoring, coupled with rating-scales already anchored conceptually. It is necessary for the rating-scales to, at the very least, be *conceptually*-anchored by the researcher so that (a) rating-scales are measuring the same concept and (b) comparisons between respondents can be made. However, the conceptual meaning can be enhanced as the aforementioned studies have shown. The interaction between these three methods of anchoring, is reflected upon. Allowing a respondent to personalise the endpoints of an already *conceptually*-anchored rating-scale, would naturally precede them *verbally*-anchoring the endpoints of the rating-scale. Should the rating-scale then possess personally meaningful verbal endpoints, this might then put them in a position whereby they could anchor their rating-scale *numerically* (i.e. define the number of categories they wish to have).

When considering the age-old problem of ‘optimal rating-scale length’, allowing each respondent to individually anchor their own rating-scale length to rate a construct of interest, could result in a rating-scale that (a) possesses personally-meaningful response categories (response categories that are closely mapped to their own ideal categories of judgment), and (b) is appropriate to respondents’ individual characteristics. It is posited that this could result in measurement reflecting respondents’ true position more accurately; that is, result in more valid measures (Viswanathan et al., 2004).

The potential feasibility of using individualised rating-scales to minimise measurement error has been considered above. However, before this idea could be tested, it was necessary to determine the feasibility of such a technique and develop a working version that would allow survey respondents to independently define their ideal rating-scales. Due to the dynamic nature of what is being proposed, it was clear that the use of technology was likely to offer the most pragmatic way forward.

### **2.7.1 The use of computer technology**

*“No one would dispute the value of a computer aid to assist the designer of surveys and questionnaires.”* (Graesser et al., 1999: 199)

As highlighted earlier, technological developments and internet penetration has provided researchers with an ever increasing amount of flexibility. As such, online research is on the increase and researchers are trying to innovate to remain competitive and current. Whilst creating a dynamic way of having respondents individualise a rating-scale, even if considered previously, may not have been feasible in a paper-based format, technology equips us with new tools, and thus new possibilities.

When researchers have experimented with the use of individualised rating-scales in the past, it has usually been in a face-to-face interview setting. This involved the interviewer administering the process and recording the response from the respondent. In this way, large data samples would have been impossible to collect, making those methods appropriate only to smaller studies or qualitative studies. A more ambitious goal is needed, to justify the need to develop a method for respondents individualising rating-scales (verbally *and* numerically). Design improvements in surveys inherently



involve cost-error trade-off decisions (Groves, 1999: 237). Since cost is eminently measurable and understood, errors in surveys tend to be given less attention than costs unless they can be measured *quantitatively*. The framing of design decisions is quantitative in nature, and errors that are quantified receive more attention than those not quantified. In this way, the aim should be to create a technique that allows respondents to individualise rating-scales, such that it can be tested in large quantitative studies in a self-administered fashion. This would justify the utility of the method and the applicability of it.

## **2.8 Summary**

Over the last couple of decades, there has been a general shift from a behaviourist approach to survey methodology to a cognitive one, with researchers becoming increasingly interested in the active role played by the respondent during the data capture process. The rise of the Internet, the sharp growth in online research activities, and the innovations made through software, have provided a plethora of possibilities for addressing old (and new) problems in ways previously inconceivable. Traditional questionnaire design has involved the researcher making difficult choices with regards to method of measurement, often with contradictory recommendations on issues relating to rating-scale length and labelling.

The researcher's measurement choices pertaining to the rating-scale length have been shown to be linked with the introduction of error in responses, in particular through the manifestation of response bias. Response bias has a detrimental impact on data, raising questions as to the trustworthiness of conclusions drawn. Techniques to correct for response bias have been criticised, and it would seem a far more useful solution if

researchers were able to minimise the impact of response bias through measurement *design*, with the subsequent improvements to reliability and validity of scores.

It is clear that in a cross-cultural context measurement equivalence becomes a very serious issue, with translation- and metric-equivalence being adversely impacted upon by rating-scales that are inappropriate to the groups under study (as a result of underlying differences between cultural groups). It is argued that whilst individual-level differences are exaggerated when comparisons are made across cultures, substantial differences are likely to exist *between* individuals from *within* a culture.

Individual characteristics, such as gender and age, have been shown by some, to be linked with response bias. Given that manifestation of response bias has been linked to inappropriate rating-scale length, these individual characteristics were considered to have a possible relationship with respondents' ideal rating-scales. There have been contradictory conclusions pertaining to the relationship between individual traits and response bias, which makes it difficult to predict the likely relationships between traits and respondents' ideal rating-scales. As such, individual traits have been examined in an exploratory fashion.

Previous attempts at having respondents individualise rating-scales have mainly been concerned with the personalising of the fixed conceptual endpoints. It is argued that should respondents be able to verbally- and numerically-anchor their own rating-scales whilst maintaining equivalent conceptualisation of the endpoints, this would produce more valid and reliable measures. Given the dynamic nature of what is being proposed, an electronic means for creating this method was deemed to be the most pragmatic way

forward, especially considering this would enable a new technique to be tested via an online survey in a quantitative study. Should the technique work, this has obvious future uses, given the technological trend and opportunities for innovation in survey methodology. Moving forwards, the research objective is:

*To develop and test a measurement instrument capable of having respondents individualise their own rating-scales for use in online surveys.*

## **Chapter 3. Methodology**

---

### 3. Methodology

#### 3.1 Introduction

This chapter outlines the philosophical stance that guided this research, the reasoning behind the chosen methodology and subsequent methods used, and mentions appropriate literature that directed these decisions.

In brief, the study consisted of a mixed-methodology, involving five steps:

Step 1: Thirteen interviews which led to the creation and development of a paper-administered Individualised Rating-Scale Procedure (IRSP).

Round 1: Interviews 1 and 2.

Round 2: Interviews 3-7.

Round 3: Interview 8.

Round 4: Interviews 9-13.

Step 2: Incorporation of the IRSP from paper-based instructions into the creation of survey software.

Step 3: Sixteen protocol-debrief interviews which further developed the IRSP in its computer-administered form.

Step 4: An online pilot test to further tune the process before large-scale testing.

Step 5: A large-scale online survey with 1,363 participants across several universities, designed to test the IRSP.

As can be seen, steps one and three are both qualitative phases of data collection, followed by a pilot test (step four) and finally a quantitative phase of data collection (step five). Step

two involved transforming the IRSP from a paper-based process into software. The research objective affected every decision taken in terms of the type of experimental design, the sample, the measurement instruments used and the planned analysis. For this reason it is important to briefly refer back to the research objective before proceeding.

The research objective was:

*To **develop** and **test** a measurement instrument capable of having respondents individualise their own rating-scales for use in online surveys.*

### 3.2 Stance within the Research Domains

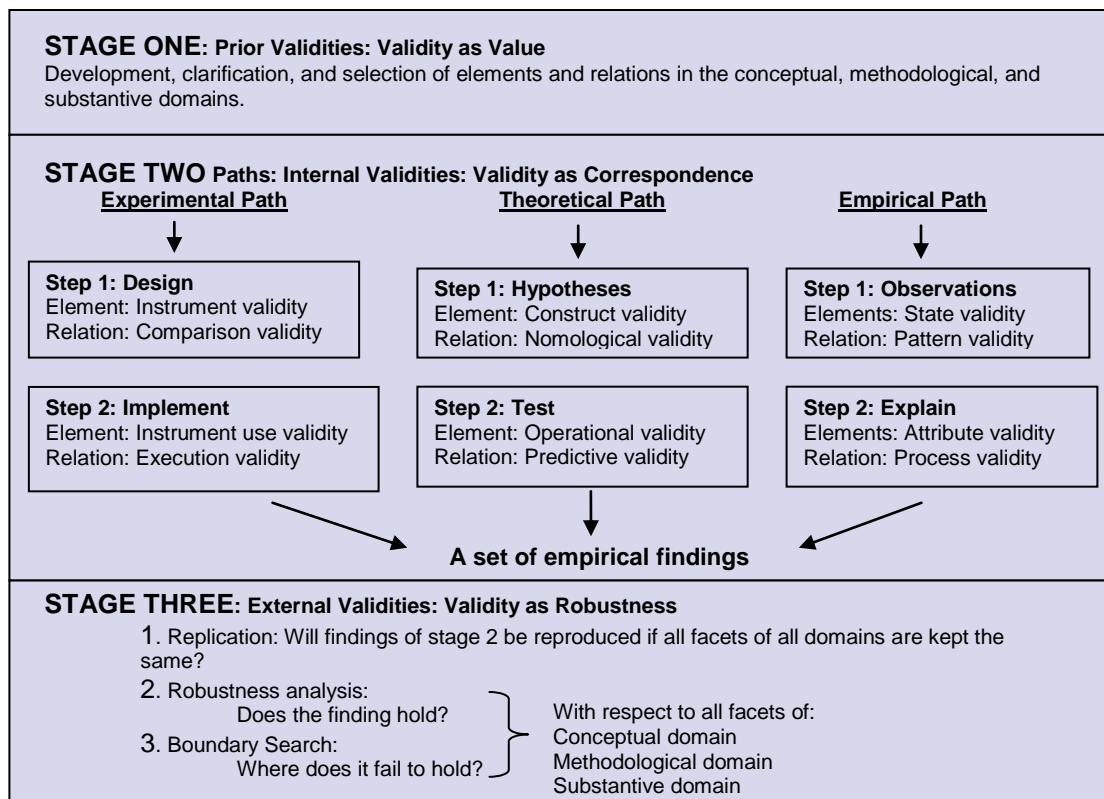
McGrath and Brinberg (1983: 117) describe the research process as, “*the identification, selection, combination, and use of elements and relations from the conceptual<sup>1</sup>, methodological<sup>2</sup>, and substantive<sup>3</sup> domains.*” The inclusion and combination of these three domains can be arranged to form three distinct research paths, which McGrath and Brinberg illustrate in their *Validity Schema*, shown in Figure 3. 1.

---

<sup>1</sup> *The conceptual domain contains elements that are **concepts**, and relations between elements that are essentially **conceptual models** about patterns of concepts (McGrath & Brinberg, 1983).*

<sup>2</sup> *The methodological domain contains elements that are **methods** – or instruments or techniques – for making observations or manipulating variables, and relations that are structures or **comparison models** for comparing (i.e. for exploring covariation and difference in) sets of observations (McGrath & Brinberg, 1983).*

<sup>3</sup> *The substantive domain contains elements that are **events** (behaviors in temporal/spatial/situational contexts) and relations that are **phenomena** (patterns of relations among events) (McGrath & Brinberg, 1983).*



**Figure 3. 1 Validity Schema as per McGrath and Brinberg (1983).**

This research addresses both Stage One and Stage Two of McGrath and Brinberg’s Validity Schema. Stage One is addressed through the *development* of an instrument capable of having respondents individualise their own rating-scales. The qualitative phase of study provided the means to *develop, clarify* and *select* elements towards the creation of a new method. This research also addresses Stage Two of their Validity Schema, by means of quantitatively *testing* the new method to establish internal validity. Within Stage Two, this project sits within the ‘Experimental path’ of the schema. In step one (design) of this path, the researcher chooses *concepts* and their relations (*conceptual domain*) and *methods* with which to test whether they hold (*methodological domain*). In step two (implementation) the researcher requires events (*substantive domain*) in order to collect data. A ‘method’ of data collection is precisely what this research sought to create. In other words, it is based heavily

within the methodological domain in step one (of Stage Two), as the objective was to develop a new measurement instrument. Due to the preliminary and experimental nature of this study, the focus was only concerned with demonstrating *internal* validity and *not external* validity (Stage Three in the Validity Schema). It is important to note that the pursuit of external validity is beyond the scope of what this study set out to achieve. Bound by the limitations and requirements of this study, there was only one clear methodological route through which to go, and the pragmatist philosophy underpinned the approach taken.

In adopting a pragmatic approach, knowledge claims arise out of actions, situations and consequences (Creswell, 2003) rather than antecedent conditions (as in post-positivism). The focal point of this research was the problem (and not the ‘methods’); *how to have respondents individualise their own rating-scales*. This philosophy allowed for the inclusion of any method (qualitative or quantitative) deemed to aid in the achievement of the research objective. Many authors have advocated for the use of pluralistic approaches to derive knowledge about the problem (Cherryholmes, 1992, Tashakkori and Teddlie, 1998).

Creswell (2003) highlights three main elements of research enquiry which inform how the entire research process is undertaken; alternative knowledge claims, strategies of inquiry and methods. These areas address the philosophical approach taken, the methodology adopted and the methods used.



### 3.3 Philosophical Perspective

According to classical pragmatists, no object or concept possesses inherent validity or importance; its significance lies only in the practical effects resulting from its use or application. The ‘truth’ of an idea or object, therefore, can be measured by empirical investigation of its ‘usefulness’. This philosophical perspective was particularly appropriate, given that this research project was heavily based within the methodological domain of research, already mentioned.

Rockwell (2004: 8) asserts that “a genuinely pragmatist view of philosophy will ultimately grant a measure of epistemic virtue to any system of thought that serves a human need”, which in this case, is the need to communicate one’s cognitions so that the receiver (the researcher) obtains a more accurate understanding of the individual’s relative standpoint.

Being that this research is within the area of business (albeit the humanist side), it is worth raising a further point that supports the philosophical approach taken, by mentioning Rupert Lodge’s thoughts on philosophies in business. He claims that,

*“Realism is objective, impersonal, quantitative, mechanistic, “contemplative”, rigid and tending towards a static program of long-range planning. It identifies business motive with the cynical selfishness recognized in the “dismal science”...Pragmatism is fluid, experimental trial-and-error, dedicated to short term goals and specific problems, and fixed with a solid eye on the human-social aspects of business life.”* (As cited in Long, 1946: 301)

Although this is a somewhat crude distinction between the realist and pragmatist philosophies, it raises a strong point. Here, Lodge argues that realism overlooks the human and morale values that support successful business, and that pragmatism possesses the humanitarian and social outlook that solves problems of a more socio-interactive context. In this way, pragmatism served this particular research problem very well, as it allowed for creativity and adaptability to changing conditions and supported a mixed-methods approach to solving problems.

The use of researcher-defined rating-scales (such as the semantic differential and Likert rating-scales), resulting from an over-emphasis on the quantitative measurement of attitudinal constructs, comes from a more classical empiricist epistemology. Under pragmatism, the previous ‘truths’ with regards to researcher-defined (e.g. Likert) rating-scales, are rejected, for they bring inherent problems and response bias to data quality. This research raised a new question; whether rating-scales should be respondent-specific and tailored, as such, to reflect the span of cognitions of *every* respondent. It can be argued that this *is* possible in a world where we now have computers, the Internet, and dynamic programs that can adapt the nature of a routine at the individual level. Thus, our ‘external reality’ has changed. As such, the techniques that the research community can now create would not have been possible one hundred years ago, when classical measurement theory was developed. In this way, a previous ‘truth’ can be improved upon, and refined in order to adapt to a changed ‘external reality’.

### 3.4 Research Design

Considering that the purpose of the research was to develop *and* test a measurement instrument capable of having respondents individualise their own rating-scales, this could not have been achieved by adopting a purely qualitative or quantitative methodology. If the intention were to simply explore the *development* of the measurement instrument, then a purely qualitative methodology would have sufficed.

Including a qualitative phase in the research design in order to *develop* the measurement instrument, meant that elements from the grounded theory approach were appropriate. In this context, the task was to derive a measurement process (method) grounded in the views of participants in the study (i.e. the development of a process that allows for respondent-defined rating-scales, would be guided by the participants' responses), that allowed for a process of development and refinement of the technique (as well as testing its feasibility). It was clear that this would involve using multiple rounds of data collection and analysis (concurrently), and the refinement and interrelationship of categories of information (Strauss and Corbin, 1998). Given that there was no pre-conceived schema of how to have respondents self-define a rating-scale (nor was the feasibility known), it was evident that theories would need to emerge directly from the data. The data would need to offer insight, enhance understanding and provide a meaningful guide to action, which is how Strauss and Corbin (1998) define grounded theory.

Whilst elements of the grounded theory approach assisted the qualitative research phase, a quantitative phase of study was vital if the second part of the research objective was to be

achieved; *test* a measurement instrument. The validity and reliability of a newly developed measurement instrument needed to be examined in order to assess its ‘usefulness’ to the research community. This could not be assessed without the inclusion of a quantitative phase of study. It was quite clear that a mixed-methods strategy was necessary in order to completely fulfil the research objective

Within the mixed-methods approach, three strategies have been frequently referred to and are the most commonly used in real world research; sequential, concurrent and transformative (Creswell, 2003, Tashakkori and Teddlie, 1998). Within this research study, the sequential procedure was quite clearly the most appropriate as it assumes that the researcher seeks to elaborate on or expand the findings of one method with another method (Creswell, 2003). The sequential *exploratory* strategy is characterised by an initial phase of *qualitative* data collection and analysis, which is subsequently followed by a phase of *quantitative* data collection and analysis. The findings of these two phases are then integrated during the interpretation phase Figure 3. 2. The requirements of this study were best met by adopting a sequential *exploratory* strategy. The rationale for adopting this approach was that firstly the requirement was to explore participants’ views in order to use this information to develop a measurement instrument, and secondly a test of the instrument on a larger sample was needed.

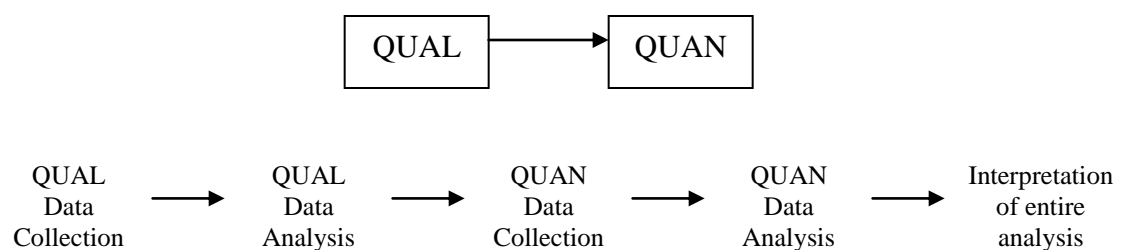


Figure 3. 2 Adapted from Creswell (2003: 213).

At the most basic level, the purpose of this strategy is to use quantitative data and results to assist in the interpretation of qualitative findings. This design is often advocated as the most appropriate to use when a researcher wishes to develop and test an instrument (Creswell, 1999), or when testing elements of an emergent theory resulting from the qualitative phase, given that it can be used to generalise qualitative findings to different samples (Morgan, 1998). It is clear that this approach is ‘useful’ to researchers who not only wish to explore a phenomenon but also wish to expand on the qualitative findings, making it especially advantageous when developing a new measurement instrument.

## **3.5 Methods**

### **3.5.1 Introduction**

A rudimentary method for having respondents self-define a rating-scale was developed, the details for which are outlined in the chapter titled ‘The Development of the Individualised Rating-Scale Procedure (IRSP)’. The rudimentary Individualised Rating-Scale Procedure (IRSP) was used in the first stage of qualitative data collection. Given that data collection and analysis occurred concurrently in the qualitative phase of this research project, it is not possible to provide a full explanation of the data collection methods used without also discussing the analysis and subsequent findings that led to iterative modification of the method. In order to keep discussions about analysis and findings out of this chapter, the entire qualitative development phase merited a chapter of its own. As such, only a brief overview of the qualitative methods used is given in this chapter. However, a detailed account of how the method was modified in stages, guided by concurrent data analysis, is

given in Chapter 4 ‘The Development of the Individualised Rating-Scale Procedure (IRSP)’.

### 3.5.2 Qualitative Phase – Feasibility test and development of the IRSP

Stage one was a qualitative exploration needed to fulfil the following research requirements:

*To test the feasibility of whether respondents can individualise a rating-scale, such that (a) they completely understand the nature of the task, (b) they can define and use it with ease, and (c) all the intervals on the rating-scale are personally meaningful<sup>4</sup> (i.e. appear to closely map their ideal rating-scales).*

Due to the exploratory nature of this phase, the most appropriate way to gain a rich bank of qualitative information was to use in-depth interviews. Focus groups for example would have been unsuitable as the phenomenon under study is the *specific feedback unique* to the individual, in order to gain insights into person-specific ways of interpreting instructions. Also, if the exercise were found to be feasible (the process of individualising a rating-scale), each respondent would have to carry it out individually (and not immersed in the views of others). Moreover, the development of the technique required that a respondent be *observed* carrying out an exercise (i.e. a sequence of instructions) and then be probed for their experience with it. This clearly demanded a one-to-one in-depth interview setting in order to achieve the research requirements of this phase.

---

<sup>4</sup> When the term ‘meaningful’ is used in this context, it refers to the distinctness of each rating-scale interval for the individual respondent, such that it possesses a meaning that is distinct from the adjacent intervals. By this definition, an individualised rating-scale that is identical to a respondent’s ideal rating-scale would be extremely *meaningful* to them.

Given the lack of pre-existing knowledge as to how an individualised rating-scale might be accomplished, elements from the grounded theory approach were appropriate for use here (Pidgeon, 1996). The planned interviews were not, however, *unstructured*, but were *semi-structured*. A ‘rudimentary’ IRSP developed for interviewees to experiment with was created, and specific questions were *planned* (which related to the participants’ use of the rudimentary IRSP). In order to have a practical conceptual basis to begin working from, it was important to consider the characteristics of interval rating-scales (e.g. equidistant intervals), and consider these characteristics in the context of an IRSP. This included considering the specific question of how to standardise scores from IRSs so that analysis could take place. From this, a ‘rudimentary’ IRSP could be grounded in key conceptual assumptions (this is outlined in Chapter 4, section 4.2.1). From this ‘rudimentary’ IRSP, new themes emerged and these were explored in the subsequent discussions. As such, maximum flexibility was needed when generating categories from the data (i.e. when coding the transcribed interviews). This creative process fully utilised the interpretive abilities of the researcher, consistent with the preliminary stages of the grounded theory approach (Pidgeon, 1996). The two main analytical practicalities that shape the methodological position which differentiates the grounded theory approach from traditional content analysis or other types of thematic analysis (Glaser and Strauss, 1967), are the method of *constant comparison* and the use of *theoretical sampling*. Both are supported primarily as a means of generating theory, as well as of building conceptual and theoretical depth of analysis and were used in this study (Pidgeon, 1996). The method of *constant comparison* involved the task of continual sifting and comparison of elements at each stage

of the qualitative data collection. In other words, there was a continuous interplay between data collection and analysis. By making such comparisons the researcher is sensitised to similarities and differences as a part of the exploration of the full range and complexity of a corpus of data, and these are used to promote conceptual and theoretical development (Pidgeon, 1996). It was necessary for the interview data to be collected and analysed in this way so that the IRSP could be developed.

#### 3.5.2.1 *Sample frame*

As per the requirements of the quantitative phase of the study (discussed later in this chapter), designed to *test* the technique, the qualitative phase needed to draw from the same sample frame for the *development* of the technique; the university student population.

#### 3.5.2.2 *Sample Size*

Given that this particular phase needed to be an exploration into the feasibility and *development* of an individualised rating-scale, and was not intended to encompass external validity, the sample size did not have to be large. It needed to be only as large as the number of 'iterations' required. In other words, sample units could continue being recruited until feasibility was established and continued improvements could be made to the exercise. However, the sample also needed to be large enough to include students from varied disciplines and both genders. There were two phases of pure qualitative data collection; stage one, which consisted of thirteen interviews informing in the development of the IRSP; stage three, which occurred after the IRSP was transformed from paper to software and was tested through sixteen protocol-debrief interviews.



### 3.5.2.3 *Sampling Method*

A type of purposeful sampling was employed. Theoretical and purposeful sampling are often confused (Tuckett, 2004), insofar as both involve a more clearly defined purpose than that involved in selecting a convenience sample. However, with purposeful sampling the sampling criteria are developed in advance of the study (and the sample does not change throughout the study), whereas with theoretical sampling, the criteria for sampling emerge along with the study itself (Koerber and McMichael, 2008). Given that prior to data collection, the purpose here was to ensure that a variety of participants from the eventual quantitative sample frame (i.e. student population) be included, this is a type of purposive sampling. It was important to include students from a range of subject backgrounds and across varying stages of study, to ensure that development of the IRSP was aided by the experiences of different ‘types of student’. More specifically this type of purposive sampling has been referred to as maximum variation (Campbell, 1999). The premise of maximum variation would be to seek to include people who represent the widest variety of perspectives possible within the range. Whilst this form of sampling is susceptible to the weaknesses associated with non-probability sampling, it was appropriate to – and fulfilled the requirements of – the qualitative development phase.

However, the fact that not all the participants have been purposefully chosen in advance, and that the number of participants included depended on the themes that emerged from the data collected, it was clear that the line between purposeful sampling and theoretical

sampling was blurred here. In cases such as these, Coyne (1997) suggested that a more accurate term for this sampling technique might be “analysis driven purposeful sampling” (as cited in Koerber and McMichael, 2008), which more fittingly describes the route that was taken.

#### *3.5.2.4 Setting*

Participants were approached in person, and were invited to participate in the interview by being asked if they could spare twenty minutes of their time. The interview settings were carefully chosen and were usually sites such as a university café or a similar public area. This allowed for a degree of background noise to make the respondent feel comfortable and relaxed, reducing a ‘formal’ atmosphere and encouraging good rapport. However, care was taken to ensure that background activity did not act as a distraction to respondents, and would not interfere with the tape recording.

#### *3.5.2.5 Data Capture*

Interviews 3-13 from stage one and all protocol-debrief interviews from stage three, were tape-recorded (permission granted in all cases) and then transcribed. A technical problem meant that Interviews 1 and 2 from round one could not be tape-recorded. In addition, an interview protocol was used for recording information. It included a heading, interviewer questions/prompts, probes to follow key questions and a space for recording comments, observations and reflective notes.

In stage one, the interviews provided valuable information as to the clarity of the experimental techniques. Participants were encouraged to discuss any problems or ambiguities with the instructions, as well as provide insights into how one might improve them further. Many probing questions (both unstructured and semi-structured) enabled the exploration of every visible avenue that may have further aided in the development of the Individualised Rating-Scale Procedure as a working measurement instrument. The central questions that guided the points covered in the interviews are detailed in Chapter 4, entitled “The Development of the Individualised Rating-Scale Procedure”.

Stage three of the project sought to explore which of two (similar) emerging paths for the final Individualised Rating-Scale Procedure was most effective (i.e. had respondents produce more meaningful rating-scales and was easiest to carry out). This stage was executed in the form of a verbal-protocol retrospective-debrief setup which had respondents carrying out a computer-administered survey using either version one of the IRSP (IRSPv1) or version two (IRSPv2).

In stage one, interviewees’ age, gender, degree scheme and whether or not they suffer from dyslexia, was noted. In stage three these demographic details were obtained as a routine part of the electronic survey. This was important in order to include respondents with varying degrees of education level within the sample (for example by including first year and final year students). This was to ensure that the instructions were being understood by a wider age group of students. Asking what degree scheme they were following, meant that it was possible to check whether students of both arts and sciences subjects interpreted the

instructions in the same way. It was important to keep a record of subjects' gender so that a roughly equal split between the number of females and males was maintained in the sample (in the event that males and females interpreted/conducted the exercise differently). Aside from the obvious information benefits from these questions, it was also a means of establishing rapport with the respondents, by asking them in an informal way how old they were and what they were studying. The question addressing dyslexia was 'useful' in order to see whether those with dyslexia had any problems understanding the instruction wording/phrasing. Given it is a more sensitive one, this question was usually posed at the end of an interview.

The themes and specific statements from participants in this first stage of data collection were analysed concurrently and indicated that the technique was feasible. As such, they were used to further guide its development. After Interviews 1 and 2, developments were made and the exercise was amended. This new exercise was tested on Interviewees 3-7. The data from Interviews 3-7 were analysed and informed further development of the exercise, which was subsequently used in Interview 8. An additional modification was made after Interview 8, and the subsequent exercise was used for Interviews 9-13. The technique was further transformed and two computer-administered versions of it were tested in stage three of the project, Interviews 1-16. As mentioned previously, a detailed account of this development is outlined in Chapter 4.

Stage 4 was a pilot test of the IRSP survey, and came before it was tested in a large scale quantitative study. The main objective of this stage was to provide a live test of the IRSP

survey in what was likely to be a similar response environment to that of the main phase of data collection. It permitted a final comparison of the two emerged versions of the IRSP (IRSPv1 and IRSPv2); it provided an opportunity to make additional checks on the survey software performance; it meant that ‘problem items’ could be noted, for later triangulation with quantitative analysis. Respondents enrolled on an MBA module were approached, permission being granted by the lecturer, for inclusion in the pilot study. The class was sent an email containing links to both versions of the IRSP survey. The variables measured by the surveys were the same as in the previous stage. MBA students were assigned to one of several lab classes as part of their module, and different lab classes were assigned to either IRSPv1 or IRSPv2. Instructions were given, observations were made, and problems were noted. Final modifications were made to the IRSP survey before the large scale survey was rolled out. Chapter 4 provides a very detailed account of the pilot test. It outlines the reasons for the decisions taken, explains in detail how the test was executed, and presents the findings.

### **3.5.3 Quantitative Phase**

Once developed, the Individualised rating-scale procedure (IRSP) was used in a quantitative phase. The overall objective of this phase was;

*To test a measurement instrument capable of having respondents individualise their own rating-scales.*

### 3.5.3.1 *Selection of a Researcher-Defined Rating-Scale*

To test the IRSP in terms of its measurement ability, to see how it performed against an existing researcher-defined rating-scale, the Likert rating-scale was chosen for this comparison. In brief, there were two key reasons which led to the Likert rating-scale's selection; it already has a large body of literature addressing its validity and reliability, and it is widely used (many constructs in surveys are made up of items that are measured using Likert rating-scales). For these reasons, it was considered useful to compare the IRSP against the Likert rating-scale (Likert, 1932).

### 3.5.3.2 *Method*

Details of the development of the IRSP can be found in Chapter 4. The IRSP is a computer-administered survey tool. As such, the *method* of data collection required a computer-administered platform. This phase of data collection showed that IRSPv2 performed slightly better than IRSPv1, and so IRSPv2 was used as the operational version of the IRSP.

The additional advantages of computer-administered data collection included: access to a larger pool of respondents (via electronic contact and a web-based questionnaire); automatic data capture and storage; time savings associated with this method of data capture; and replicating a likely research setting where the IRSP might be used.

### 3.5.3.3 Design

The quantitative test was concerned with a form of method bias (van de Vijver and Leung, 1997), in that the two *measurement methods* were the *independent variables* in this experiment. This required a multi-group experimental design to compare the measurement properties (dependent variable) of one measurement method (IRSP) with another (Likert-type rating-scales) (Robson, 2002, Campbell and Stanley, 1966). Put simply, this next test needed to be able to compare the measurement properties of the IRSP with those of Likert-type rating-scales. This required that the experimental design involve repeated measures, in that one group of respondents needed to complete a survey using IRSP (hereafter referred to as  $X_I$ ) in time period 1 (hereafter referred to as T1), and would complete the same survey, only, using Likert-type rating-scales (hereafter referred to as  $X_L$ ) in time period 2 (hereafter referred to as T2).

In order to counteract order-effects in terms of its impact to internal validity, a second group would need to be given the treatments in reverse order;  $X_L$  in T1 followed by  $X_I$  in T2. For example, should a respondent who uses the IRSP in T1 become aware of how they are better able to represent their span of judgement, this might alter how they subsequently use the Likert-type rating-scale (LTRS) in T2. Having a second test group with the treatment-order reversed would help to reduce this order effect.

In order to determine the test-retest reliability of both the new IRSP and the Likert-type rating-scale, there would also need to be two additional test groups; one receiving  $X_L$  in T1 and again in T2, and the other receiving  $X_I$  in T1 and again in T2.

Randomly allocating respondents (R) into these four test groups would increase the internal and external validity of the experiment by making it a true experimental design. Typical experimental design notation uses an  $O_i$  when referring to a process of measurement, and  $X_i$  when referring to the exposure of a group to an experimental variable or event. Given the experimental variables have already been defined above as  $X_L$  and  $X_I$ , it is also clear that this experiment is not quite the norm, in that the *exposure* to treatment ( $X_i$ ) and the *measurement* ( $O_i$ ) occur simultaneously. In this way, this design could not be defined as a classic pretest-posttest, but more like a test-retest.

**Table 3. 1 True multi-group experimental design**

Test Group	T1	T2
TG <sub>1</sub>	R [ $X_I O_1$ ]	[ $X_L O_2$ ]
TG <sub>2</sub>	R [ $X_I O_3$ ]	[ $X_I O_4$ ]
TG <sub>3</sub>	R [ $X_L O_5$ ]	[ $X_I O_6$ ]
TG <sub>4</sub>	R [ $X_L O_7$ ]	[ $X_L O_8$ ]

### 3.5.3.4 Validity

Below are factors that typically jeopardise internal and external validity, as described verbatim by Campbell and Stanley (1966), and were relevant for consideration when generating the above design. Outlined below each, are the considerations given to the above design for controlling these effects.



**Table 3. 2 Extraneous variables that affect internal validity**

<b>Internal Validity</b>		
<b>Extraneous Variables</b>	<b>How it manifests</b>	<b>How was this addressed</b>
History	Specific event occurring between the first and second measurement.	In the above design, <i>history</i> is controlled insofar as any potential historical events that might have produced a difference in O <sub>1</sub> -O <sub>2</sub> would, on the whole, also produce an O <sub>3</sub> -O <sub>4</sub> , O <sub>5</sub> -O <sub>6</sub> and O <sub>7</sub> -O <sub>8</sub> difference. The random assignment of students into TG <sub>i</sub> s, together with the TG <sub>1-4</sub> being measured simultaneously, reduces the likelihood of <i>history</i> affecting validity.
Maturation	Processes within the respondents operating as a function of the passage of time per se.	In the above design, <i>maturation</i> is controlled in that it should manifest equally in both experimental and control groups.
Testing	The effects of taking a test upon the scores of a second testing.	In the above design, <i>testing</i> is controlled in that it should manifest equally in both experimental and control groups. However, the additional issue for consideration was the order-effect of the survey completed first, X <sub>L</sub> or X <sub>I</sub> . TG <sub>1</sub> being the inverse of TG <sub>3</sub> in terms of treatment-order should control for differences that would have been attributed to order-effects.
Instrumentation	Changes of the calibration of a measuring instrument may produce changes in the obtained measurement.	This is precisely the focus for this experiment, and so does not apply here in the classical way.
Selection	Biases resulting in differential <i>selection</i> of respondents for the comparison groups.	In the above design, <i>selection</i> effects would be discounted as an explanation of any differences, to the extent that respondents were randomised into test groups.
Experimental mortality	Differential loss of respondents from the comparison groups.	In order to minimise the effects of <i>experimental mortality</i> on the above experimental design several actions were implemented: <ul style="list-style-type: none"> <li>○ A combination of two survey incentives (deemed to be quite different in terms of the type of respondent each would attract) were only offered to respondents who took part in <i>both</i> T1 and T2. This was explicitly stated in all invitation-to-participate emails/links.</li> <li>○ Reminders were sent to respondents who appeared to be delayed in their return to complete the survey in T2, and a record of those needing a reminder were kept (to test for differences between those who returned with no reminder, and those who returned as the result of a reminder).</li> <li>○ A large sample frame was obtained so that the test groups were large enough in T1 to cope with <i>mortality</i> effects, leaving the test groups with an adequate amount of respondents in T2.</li> </ul>

**Table 3. 3 Extraneous variables that affect external validity**

<b>External Validity</b>		
<b>Extraneous Variables</b>	<b>How it manifests</b>	<b>How was this addressed</b>
<i>Reactive or interaction effect of testing and <math>X_i</math>.</i>	A pretest (in this context, the first test) might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected.	A 'wash out' period of between 4 -8 weeks was introduced between the testing in T1 and T2. This ensured that enough time would pass to allow respondents to forget their responses in T1. The 'wash out' period was not longer, given this could have increased the affects of <i>maturation</i> and <i>experimental mortality</i> . Minimising the affects of these variables was important to maintaining <i>internal validity</i> .
<i>Interaction effects of selection on <math>X_i</math></i>	Effects of <i>selection</i> biases and the <i>experimental variable</i> .	This project was not intended to generalise the results of this study to the general population. However, in order to be able to generalise (to some degree) the results to the British student population, several universities were included in the sample frame. They are all in geographically different areas and possess varying student demographics.
<i>Reactive effects of experimental arrangements</i>	A prominent source of unrepresentativeness is the artificiality of the experimental setting.	The respondents were recruited electronically and were therefore already in a setting (such as a library, office, or at home) in front of a computer, which is exactly where they would be in a 'real world' online survey. As such, the setting chosen for this study mimicked the 'real world' setting.

### 3.5.3.5 Sample Size

When considering how many sample units would be required for each sample, the sampling distribution of the statistics was considered. Based on the central limit theorem, when the sample size reaches around thirty, the statistical assumptions related to normality are approximated (Hair et al., 2003). Therefore, an absolute minimum of 30 respondents was required in each test group. However, a further increase in sample size was targeted in order to increase the confidence level (stemming from the Central Limit Theorem), and to gain greater precision in sample results (Lipsey, 1990). As previously mentioned,

*experimental mortality* also meant that larger numbers would need to be targeted, than were needed. Additionally, in order for there to be some flexibility as to the variables examined through statistical analysis (e.g. differences between males and females), it was an objective to make each test group as large as possible.

Additionally, more respondents were targeted for the  $X_I$  groups than  $X_L$ , in T1. This was done for two reasons: (a) to pre-empt a scenario where the experimental mortality rates were high (perhaps during a student exam period), leaving little T2 data, the need to collect as much data as possible on the *new* method ( $X_I$ ) was prioritised; (b) because the  $X_I$  takes slightly longer to complete than the  $X_L$ , it was necessary to pre-empt the possibility that experimental mortality rates might be higher for those doing  $X_I$  in T1. For these reasons, it was planned that approximately two thirds of respondents in T1 would complete  $X_I$  with the rest completing  $X_L$ . The numbers obtained for each of the experimental groups are shown in Table 3. 4.

**Table 3. 4 The sample size achieved by each test group.**

Test Group	T1	T2	Experimental mortality	No. of respondents that completed T1 <i>and</i> T2.
TG1	[ $X_I O_1$ ] N = 386	[ $X_I O_2$ ] N = 282	104	N = 282
TG2	[ $X_I O_3$ ] N = 393	[ $X_L O_4$ ] N = 297	96	N = 297
TG3	[ $X_L O_5$ ] N = 293	[ $X_I O_6$ ] N = 202	91	N = 202
TG4	[ $X_L O_7$ ] N = 291	[ $X_L O_8$ ] N = 213	78	N = 213

### 3.5.3.6 *Sample frame*

In order to ensure that the findings of this research could comment on how the use of particular measurement instruments affect survey data, the impact of extraneous variables needed to be reduced as much as possible. As such, it was important to minimise differences between sample units. To this aim, a homogenous sample was desirable, as this would contribute towards establishing internal validity of the new method. A sample frame that possessed similar demographic characteristics, such as age, level of education, and culture, would reduce the impact of extraneous variables on the experimental design. The study was concerned only with differences between the methods of measurement (the independent variables) on the data collected, and not with variation of any other kind. Bearing these issues in mind, university students were deemed the most suitable sample frame given the relative homogeneity of their level of education and age (certainly when compared to the general population). This is also a sample frame that was more readily accessible than others, augmenting the likely success of data collection. The decision was taken to use the student population as a suitable sample frame for the qualitative phase, and the subsequent quantitative phase of testing.

With regard to the actual sample frame obtained for the quantitative data collection, a great deal of networking was done in order to befriend potential ‘gatekeepers’ who could provide access to large student groups. In this endeavour, the strategy was to collect data from a spread of different universities (located in various geographical areas) and to collect data from a wide variety of students across degree disciplines. This was so that the sample could better represent the student population. Table 3. 5 shows the location of each gatekeeper,

the access offered through particular channels (entry points), and the approximate size of the sample frame for which access was offered.

**Table 3. 5 Gatekeepers secured for sample frame**

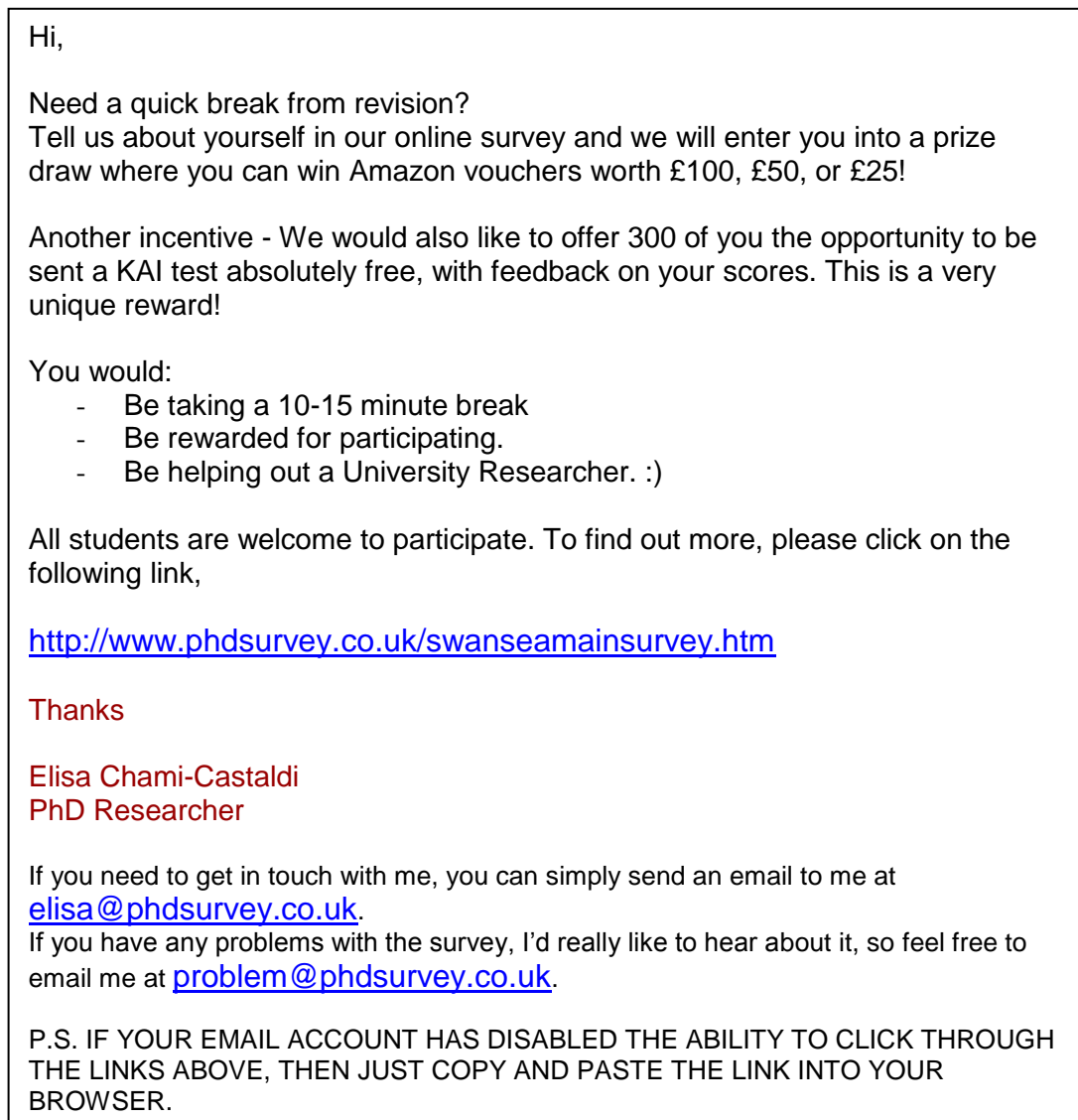
Gatekeeper	Gatekeeper Location	Access Offered	Approximate Size of Sample Frame
1	Bradford University – various gatekeepers	Survey link featured in Campus News email which is sent to all students and staff. Survey link featured on the Internal Homepage's News section. Survey link featured on the School of Management's internal web page. Forwarded an email on my behalf, to all Management School students.	13,070 students (all)
2	Leeds Metropolitan Law School	Forwarded an email on my behalf, to Law students at Leeds Met University and sent out a message through the Law portal. Placed the survey link on the intranet Law portal.	900 undergrad students, 160 postgrad law students.
3	Department of Modern Languages University of Exeter and Psychology Department	Forwarded an email on my behalf, to all undergraduate students in the modern languages department and psychology department at University of Exeter.	890 undergraduate students of Modern Languages and Psychology
4	Swansea University	Forwarded an email on my behalf, to all Swansea University students.	13,825 students (all)
5	MD of I-Graduate Ltd	Forwarded an email on my behalf to English speaking British students and International students in Canada and Australia.	3000 students in total from a broad range of universities.

All the gatekeepers and their corresponding student groups were accessed during the quantitative data collection phase (in the manner described in the table).

### 3.5.3.7 Sampling Method

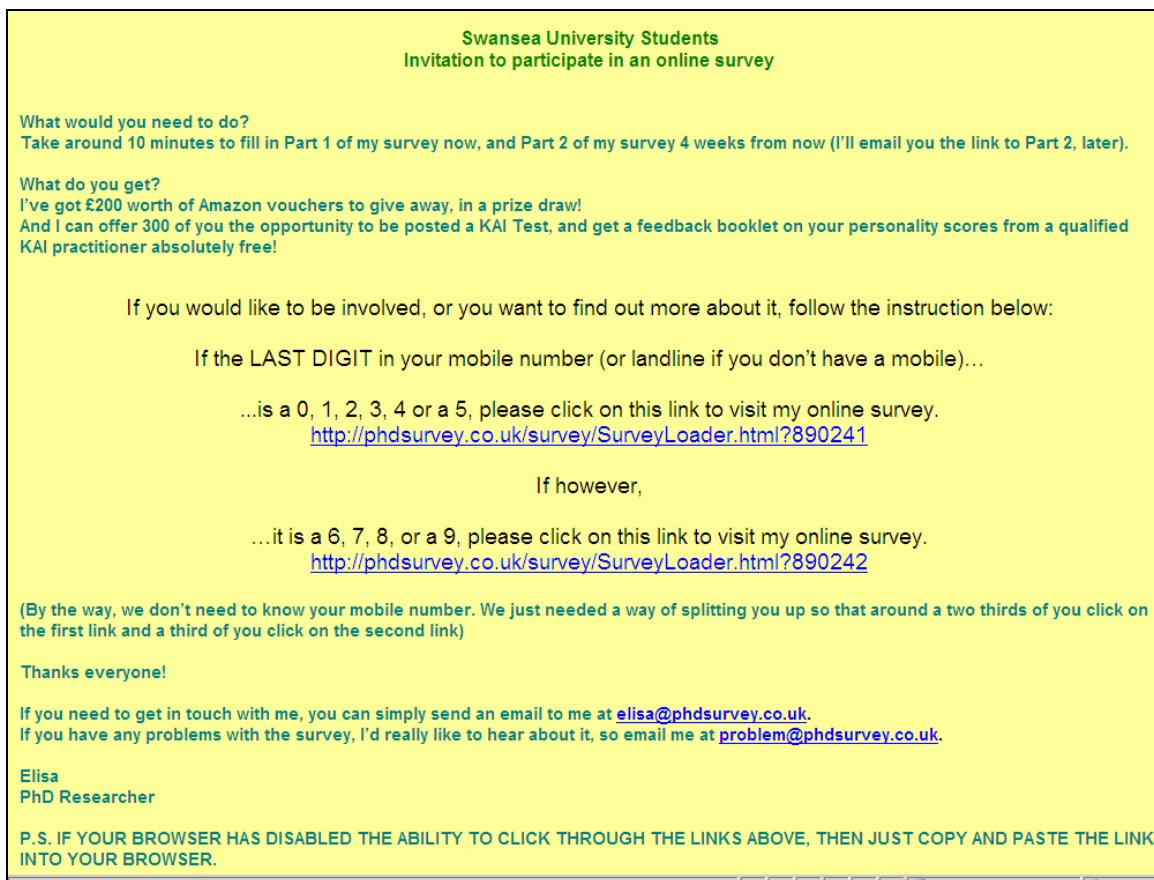
The above mentioned gatekeepers were sent out the invitation-to-participate emails so as to forward them on to the respective student sample frames. Figure 3. 3 shows the text that was included in the email, which was forwarded onto the students. In addition to these emails, the gatekeeper at Leeds Metropolitan University placed the survey invitation on their intranet web portal for Law students. Moreover, a survey invitation link was also

placed on Bradford University's main intranet page, as well as on the Management School's web page. As such, there were several survey entry points, depending on *where* the respondents had seen the survey invitation within and across universities.



**Figure 3. 3 Invitation-to-participate email sent to students at several universities.**

Clicking on the link contained in the email would take respondents to a page like the one shown in Figure 3. 4.



**Figure 3. 4 Survey welcome page assigning respondents into groups.**

A mobile telecommunications infrastructure consultant confirmed that the last digit of mobile phone numbers were uniformly spread from 0-9. He indicated that there was an equal probability of one's last digit being any of the numbers from 0-9, (i.e. mobile phone providers are not in the habit of favouring any of the numbers for the last digit).

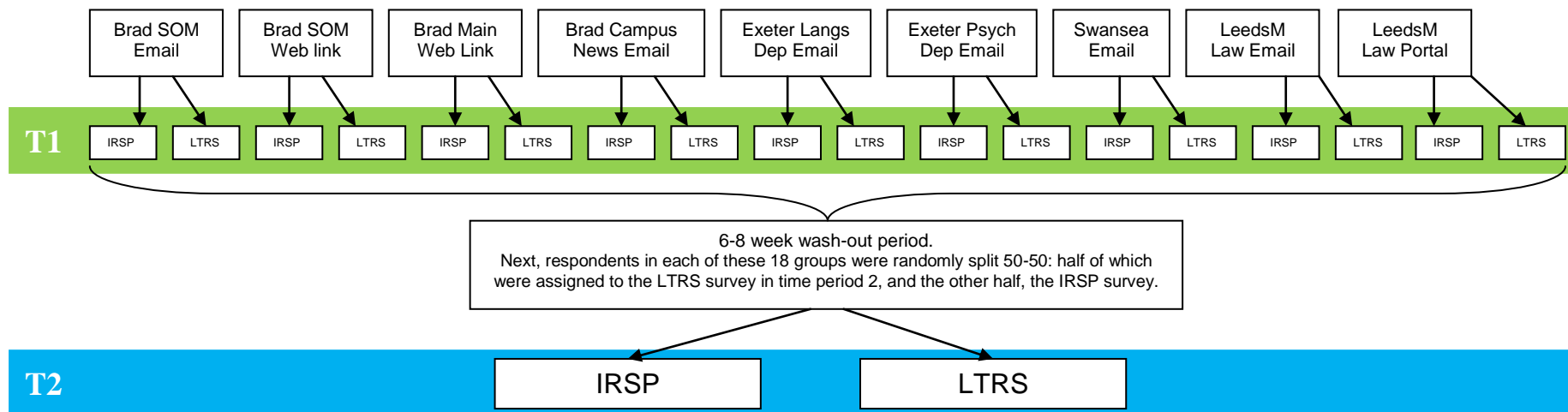
Randomly assigning respondents to either IRSP or LTRS in time period 1 (T1) was done in a manner that meant there would be a slightly higher portion assigned to the IRSP group. This was done to ensure that, should experimental mortality rates have been high, there was enough data available for the new IRSP method in T1 to perform sufficient analyses. As such, only four out of the ten possible digits directed respondents to the LTRS group in T1. As it happens, a large number of responses were obtained in T1 and mortality rates were

low. However, the mechanism, assigning marginally more sample units to the IRSP group, was a necessary measure in case data capture had not been as successful.

Previously mentioned was the fact that the IRSP survey software was able to track exactly through which entry point each respondent had reached the online survey in T1. This was very useful information, given certain types of students may have been more or less likely to view a particular entry point. For example, the more organised and conscientious students might check their departmental portal frequently, and therefore be more likely to take part in the survey than other students. Through tracking the entry point of each respondent, any bias that could have been introduced here was catered for; sample units were randomly allocated into their time period 2 (T2) test groups, by entry point. This is best illustrated in Figure 3. 5.

There were only a small handful of emails sent by respondents to [problem@phdsurvey.com](mailto:problem@phdsurvey.com). It was established that these respondents had experienced difficulty in completing the survey either because they were using a Linux operating system or because they had an old version of Adobe Flash. Although this will have meant that a small number of respondents had difficulty with completing the survey, this was unlikely to have significantly biased the data.





**Figure 3. 5 Random assignment into test groups by survey entry point.**

Respondents were offered two types of incentive. One was the opportunity to win £200 pounds worth of Amazon vouchers (1 x £100, 1 x £50, and 2 x £25). The other was the opportunity for 300 (who opted in) to be posted out Kirton Adaption Innovation (KAI) tests, with feedback on their scores offered only after their completion of the survey in T2. Offering two very different incentives was done in an effort to have survey participation appeal to differing types of students (e.g. perhaps attracting different personalities). Whilst some may have been less interested in the vouchers, they may have been more interested in the personal insight they could get from the KAI feedback. During the data collection, an additional 150 KAI tests were acquired, which meant that a total of 450 KAI tests were available to be sent out to respondents. More than this number had opted in to receive the KAI. As such, respondents were randomly chosen to receive the KAI test. A total of 403 KAI tests were returned completed. This data would provide an interesting area of further research, however it was not possible to include the KAI data within the scope of this study, due to length limitations, and the fact that it would go above and beyond the objective of this research.

#### *3.5.3.8 Setting*

Respondents were targeted for T1 participation in April 2008. This will have meant that some may have been busy preparing for exams, and therefore too busy to participate. However, it may also have meant that some students may have been checking their university emails and intranet portals more frequently (for precisely the same reason), and indeed have welcomed the procrastination afforded through participation. Respondents were invited back for the T2 re-test in June/July, which would have fallen after their exams. Whilst some students could have been on holiday, others will have had more time to respond given the exam period was over.

The likely response setting will have been either the library, at home, or at work. This is likely to closely mimic a typical online survey setting.

#### *3.5.3.9 Survey items*

Following on from the individual characteristics (both demographic in nature, and personal traits) discussed in the literature, choices over survey item inclusion stemmed from those theoretical considerations highlighted. The online survey captured respondent demographics; surname, email address, home address, student number, gender, date of birth, total years in university education, postgraduate/undergraduate status, degree subject, first language, ethnicity, and national identity. These items were necessary in order to: gain an understanding of the sample; test individual demographic characteristics for their relationship to Individualised Rating-Scales (IRSSs) defined; make demographic comparisons by groups; and exclude certain cases/groups from analyses where necessary.

The psychometric items included in the survey were: the Affective Orientation (AO) fifteen-item scale developed by Booth-Butterfield and Booth-Butterfield (1996); the eleven-item, two-factor, Personal Need for Structure scale, as per Neuberger and Newsom (1993); the Cognitive Style Indicator (CoSI) eighteen-item, three-factor scale, as per Cools and Van den Broeck (2007); and the Big Five Inventory (BFI) ten-item, five-factor scale, as per Rammstedt and John (2007). The survey also included two items measuring current 'mood', and four items designed to gain feedback from respondents on the use of the IRSP (with regard to ease, meaningfulness, attention given, and preference).

Neuberg and Newsom's (1993) Personal Need for Structure (PNS) scale was included given its focus is on cognitive structuring, which is the "creation and use of abstract mental representations" (1993: 113). As already mentioned in the literature review, respondent's abilities to use their abstract faculty has been linked with the manifestation of response styles resulting from inappropriate rating-scale length. The PNS scale's focus on cognitive structuring is therefore of interest given there could potentially be a link between this personal characteristic and one's ideal rating-scale.

Neuberg and Newsom (1993) had administered the PNS scale to six independent groups of undergraduate male and female students at American universities, with over 2,900 respondents in total. The PNS scale proposed by Neuberg and Newsom (1993) consists of 11 reflective items forming a two-factor measurement model: the extent to which people prefer to structure their lives (Desire for Structure) and the manner in which people respond when confronted with unstructured, unpredictable situations (Response to Lack of Structure). The scale was validated using a 6-point Likert rating-scale, with all six intervals labelled (strongly disagree, moderately disagree, slightly disagree, slightly agree, moderately agree, strongly agree). They applied confirmatory factor analysis (CFA) to the data and reported that this model achieved acceptable fit with adequate internal reliabilities for the overall scale and the two constituent factors (with the median Cronbach alpha =.77). They highlighted that the two factors correlated highly (with inter-factor correlations ranging from .54 - .75 across the six sample), but that this model was still a better fit than a one-factor alternate version. They argued that one would expect the two factors to be highly related. Good test-retest reliability was present for both factors (.84 and .79 respectively across a 12-week period).

Booth-Butterfield and Booth-Butterfield's (1996) Affective Orientation (AO) scale was included because it measures respondents' affective orientation as a guide to their communication and behaviour. Respondents that are high in their affect orientation might be more or less prone to defining rating-scales of a certain length. This scale measures the degree to which one uses their emotions to guide their actions and the way they communicate. For this reason, and also because Neuberg and Newsom (1993) theorised there might be a connection between one's desire for structure and one's affective experience, it was included.

The authors had administered the scale in two studies in America; one with 124 working adults enrolled in applied communication courses and the second with 148 undergraduate students. The AO scale consists of 15 reflective items which are measures of one-factor. It was validated by the authors using a 5-point Likert rating-scale, with all five intervals labelled (strongly disagree, disagree, uncertain, agree, strongly agree). The authors applied CFA to the data and reported that this model achieved acceptable fit with adequate internal reliabilities for the overall scale in both groups (with Cronbach alpha = .88 and .92), and it possessed face validity and discriminant validity. Booth-Butterfield and Booth-Butterfield indicated that their model provided a reasonably good fit with "the obtained values in both samples for the one-factor model all approach[ing] .90 which is sometimes used as a conventional criterion of good fit" (1996: 160). However, they indicated that "it was clear that there was still some unexplained variation in the data missed by the one factor model" (1996: 160) and that they had experimented with the removal of certain items but could not

identify a model that provided what they considered to be a better conceptual or empirical fit to the data.

Cools and Van den Broeck's (2007) Cognitive Style Indicator (CoSI) scale was included because of its links to the CASM movement (Cognitive Aspects of Survey Methodology). The authors define cognitive style as "the way people perceive stimuli and how they use this information to guide their behaviour" (2007: 360). Obtaining a deeper understanding of the types of rating-scales defined by respondents with differing cognitive styles, would be a valuable contribution in the CASM context. This particular scale was used because the authors were able to consolidate a myriad of cognitive style models and measures from a complex field of study, culminating in the production of a scale which measures a respondent's employment of three distinct cognitive styles: those that look for facts/data, want to know exactly the way things are, and like complex problems if they can find a clear and rational solution (knowing style); those that are characterised by a need for structure, like to organise and control, and attach importance to preparation and planning (planning style); and, those who tend to be creative, like experimentation, uncertainty and freedom, and who see problems as opportunities (creating style). An individual can possess one or a combination of all these cognitive styles to varying degrees.

After scale development was conducted through two large studies with an educated sample of the general Belgian population, the scale was validated in a study consisting of 635 MBA students from Belgian business school. The CoSI scale consists of 18 reflective items measuring the three-factors. It was validated by the authors using a 5-point Likert rating-scale, with the endpoints labelled (totally disagree and totally agree).

The authors applied CFA to the data and reported that this model achieved acceptable fit with adequate internal reliabilities for the overall scale (with Cronbach alpha =.76 for knowing style, .85 for planning style and .78 for creating style), and it possessed face validity and discriminant validity.

Given that the quantitative study was exploring whether relationships existed between individual characteristics and individualised rating-scales, it was deemed important to include a scale that measured broader characteristics and not just scores on more narrow traits (like PNS, AO and CoSI). Rammstedt and John's (2007) Big Five Inventory scale (BFI) was included because it measures the broader five facets of personality (Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness) using a short ten-item measure, which was considered to be very advantageous given typical BFI scales are much longer and would have been too long to include in this study.

This scale was developed based on a 44-item version developed by John et al. (1991). It was validated on two samples, each consisting of 726 US university students, and two German samples consisting of 457 and 376 students respectively. It was validated by the authors using a 5-point Likert rating-scale, from 1="disagree strongly" to 5="agree strongly". The authors evidenced construct validity and strong test-retest reliability with coefficients ranging between .72 - .78. However, their results suggested that given the scale's brevity, there were substantial losses in comparison to the full 44-item BFI scale. They recommend that if testing time is not limited, that the full BFI scale be used, given its psychometric advantages. However, given this survey would have been far too long should a 44-item measure have been included (and would be prone to bias associated with lengthy surveys, mentioned in the literature), the shorter ten-item measure of BFI

was included given it is suitable when uncovering facet-level relationships (John et al., 1991).

#### *3.5.3.10 Data Capture*

Survey data was backed up frequently (usually daily), in the event that external factors impacted on data collection (e.g. a server going down or website maintenance). Data capture was successful, and free of problems. Reminders were sent to respondents who had not completed the re-test within a reasonable time frame (usually around the two-week mark). This resulted in the majority, who were reminded, returning to complete the survey. Data was scanned and cleaned: the removal of duplicates (respondents who responded more than once – although there were few); the correction of spelling mistakes on the verbal labels chosen with the IRSP survey; incomplete entries being removed (which were fewer than approximately 5% total responses). Those that had incomplete entries were emailed in order to ascertain the reason for the incomplete entry. This was done to ascertain whether there had been a problem with the survey itself. In most cases, the problem had occurred due to a temporary loss of internet connection (which the survey software was not robust to). These respondents were invited to try again when their internet connection was more stable. A small number (approximately ten) had indicated they had lost interest half way through.

#### *3.5.3.11 Planned Analysis*

The planned analyses were intended to test for whether the IRSP data was able to replicate the psychometric measurement models in a similar fashion to that of the LTRS data (designed for those very psychometric scales). This was done via a comparison of groups for T1 data (IRSP vs. LTRS), through the application of confirmatory factor



analysis in Amos 7. This process progressed from loose cross-validation of the measurement models between the groups, to tight cross-validation. This was planned in order to assess metric and scalar equivalence between the methods and the measurement models produced through them. In addition, convergent and discriminant validity of the measurement models were examined using the Campbell and Fiske's multitrait-multimethod matrix. The test-retest reliability of the two groups were also compared, through correlations between the psychometric factor scores in T1 with T2. Relationships between individual characteristics and individualised rating-scales chosen were explored, through a series of parametric and non-parametric tests. Finally, respondents' feedback and their preferences over measurement instrument were examined through descriptive statistics.

#### **3.5.4 Ethics**

As per the University of Bradford's Code of Research Ethics, appropriate arrangements were made to obtain informed consent from each participant and respondents' data were securely protected and used appropriately. To make sure that as a researcher, actions taken were appropriate and in accordance with ethical standards, the Market Research Society's Code of Conduct was consulted.

### **3.6 Summary**

This chapter began with an outline of the philosophical stance that guided this research. Next, the reasoning behind the chosen methodology was explained along with the specific methods chosen, sampling considerations, and the approach planned for data analyses.

Given that the research objective was concerned with *developing* and *testing* a measurement instrument, a mixed-methodology was chosen as the most appropriate approach. The research project was divided into five steps: a qualitative phase of development; design and build of the IRSP software; further refinement and development; a pilot test; and finally, a large quantitative phase of online testing.

Analysis driven purposeful sampling was employed during the qualitative development phase so as to induce maximum variation in the sample pool so that development could address the widest variety of perspectives. Analyses during this qualitative phase were iterative and influenced additional sampling requirements as well as development of the IRSP method. A multi-group experimental design was chosen to execute the quantitative phase of testing. This was done so that the test-retest reliability and validity of the new IRSP method could be compared with an existing method (i.e. LTRS). It was decided that the sample needed to be homogenous across variables that could potentially impact on response bias manifestation and thus internal validity of the experiment. The student population was considered to be appropriate. Respondents were randomly assigned into experimental groups. Structural equation modelling (SEM) was used to investigate whether the IRSP data was able to replicate the psychometric measurement models in a similar fashion to that of the LTRS data (by which those models were originally validated). Internal validity and test-retest reliability was also examined through SEM.

## **Chapter 4. The Development of the Individualised Rating-Scale Procedure**

---

## **4. The Development of the Individualised Rating-Scale**

### **Procedure (IRSP)**

#### **4.1 Introduction**

This chapter covers the four stages of the development of a method for respondents to self-define a rating-scale, hereafter referred to as the Individualised Rating-Scale Procedure (IRSP):

Stage 1: Thirteen interviews led to the creation of a paper-administered IRSP.

These thirteen interviews were subdivided into four rounds. A new round was initiated when there was enough data obtained from the previous interview(s) to inform a significant development in the IRSP. The IRSP was then modified and tested further. As such, the IRSP was adjusted between each round, and this process of refinement is what differentiates the rounds.

Stage 2: Creation of survey software for the IRSP.

Stage 3: Further development and testing of the computer-administered IRSP, through sixteen verbal protocol-retrospective debrief interviews.

Stage 4: An online pilot to further tune two versions of the IRSP before choosing to take one forwards onto large-scale testing.

The qualitative insights gained are discussed at every stage of the development, as appropriate.

#### **4.2 Stage 1: Foundations for the Individualised Rating-Scale**

##### **Procedure (IRSP)**

Stage one was a qualitative exploration to fulfil the following research requirements:

*To investigate whether respondents can individualise a rating-scale, such that: (a) they completely understand the nature of the task; (b) they can define and use it with ease; (c) all the intervals on the rating-scale are meaningful.*

#### **4.2.1 Assumptions underlying the development of the ‘rudimentary’ IRSP**

The literature review highlighted three ways in which a respondent can individualise a rating-scale:

- Verbally anchor the rating-scale endpoints;
- Numerically anchor the rating-scale endpoints (i.e. defining the number of response categories available to use);
- Conceptually anchor the rating-scale endpoints (although in previous studies these have always been fixed by the researcher, with the respondent sometimes being asked to personalise the concepts by imagining a specific state/scenario).

These methods have previously been used in isolation, however, there is a lack of research that investigates the use of more than one simultaneously. This research sought to explore whether respondents could, in fact, individualise a rating-scale, verbally *and* numerically. The conceptual anchor needed to remain constant across respondents so that inter-respondent comparisons remained possible. It is possible that two respondents’ absolute extremes of opinion could differ (i.e., respondent A could have a more extreme level of agreement than respondent B). However, an implicit assumption underlying inter-respondent comparisons of standard LTRSs is that the conceptual end-points of the scale are equivalent across respondents. Responses on existing LTRSs implicitly measure how close each respondent’s opinion is to their most extreme opinion. It is beyond the scope of this study to investigate whether this assumption is valid. As such it is assumed that the end-points of the IRSs are conceptually equivalent.

Initially, there needed to be a ‘starting point’; a technique to use as a reference point to begin the qualitative exploration and investigate the feasibility of respondent-defined rating-scales. Respondents could provide insights into the cognitive processes they experienced as a result of doing that activity. Thus, refinement would involve tuning any processes that were shown to be *useful* (pragmatist thinking) until a working measurement instrument, grounded in the views of the participants, was developed.

Reflecting upon the components possessed by classic researcher-defined scales (e.g. Likert-type rating-scales), and the subsequent data analysis generated by collecting data using these rating-scales, a measurement instrument needs to have certain components to make data analysis possible (e.g. assumed equal intervals, and conceptually equivalent endpoints). Knowledge of the characteristics of a measurement technique that allow sound data analyses, is essential when developing a new method of measurement. In this context, it was important to consider what to incorporate into a new dynamic process for having respondents individualise rating-scales, in order for the measurement method to remain *useful* for researchers whilst also being analysable. Discussions with experts were used as a means of assessing the components deemed necessary (or *useful*) to include in the formation of the ‘starting point’ for the qualitative exploration. This ranged from a series of in-depth discussions with academic researchers and more general feedback from the academic community at conferences.

Further rating-scale characteristics, as well as the resultant operational decisions, are outlined below:

a. Representing the “range”.

Visually representing the range of possible responses on a rating-scale can be achieved vertically (e.g. Nugent, 2004) or horizontally. With typical interval rating-scales, researchers’ use a horizontal line. This may have occurred because it allows multiple response rating-scales to be fitted on a typical portrait questionnaire page. Nevertheless, as Western respondents typically read text horizontally, response categories presented horizontally also mirror the common reading format. It is beyond the scope of this study to test whether this familiarity results in more effective or efficient information processing when response categories are presented horizontally. However, in this study it is assumed that a respondent’s absolute span of judgment is best represented by a horizontal line. This horizontal line is provided as a pictorial aid to consider their absolute span of judgment. This helps elicit a respondent’s cognitive span on an issue, and provides a familiar pictorial medium to conceptualise their span as a continuum. This pictorial representation was considered to be useful and appropriate, and was therefore included as part of the initial IRSP.

b. Anchoring specific values.

Anchoring occurs, as a minimum, at the endpoints of the rating-scale and can be conceptual, verbal and/or numerical (as well as pictorial). Conceptual anchors are assumed to be equivalent at the most extreme level of agreement (or disagreement) for the reasons detailed previously. Verbal and numerical anchoring communicates some sense of maximum/minimum meaning to the endpoints (e.g. strongly agree = 7, strongly disagree = 1). Verbal labels are interpreted subjectively. The precision (i.e. consensus as to meaning) of most adverbs and adjectives used to express frequency/degree/meaning, is inadequate (Nakao and Axelrod, 1983). Worthy of note, is that this has been shown to

be a serious problem in single-country research where respondents speak the same first-language. For example, for respondent A, the term ‘strongly agree’ might constitute as complete an agreement with something as is possible for them. However to respondent B, the very same verbal label might not encompass their absolute stance on the agreement continuum, and could be interpreted to be several degrees short of their ‘absolute’. This creates a problem for researchers; one cannot be sure that verbal qualifiers are calibrated between *all* respondents, and are being interpreted consistently. As such, an exploration of methods for individualised verbal anchors needed to be included in the qualitative development of the IRSP.

Given that this research proposes that respondents be involved in the process of lengthening/shortening rating-scale length (i.e. the number of intervals), there needed to be an exploration of how best to instruct respondents to numerically anchor their rating-scale intervals (within the qualitative development process).

From the literature, numeric values are sometimes used to disambiguate the meaning of verbal labels. This results in respondent-specific interpretations and different ideal rating-scale anchors (Schwarz et al., 1991a). Consequently, respondents first need to have a conceptual understanding of their positions, then verbalise them (i.e. verbally anchor their cognitive span of judgment), *before* attaching numerical values (i.e. before determining the number of intervals they have). For example, hypothetically, if told that you are currently feeling ‘neutral’ about an issue (i.e. absence of ‘agreement’ or ‘disagreement’), but if presented with an issue that you agree with as much as you possibly could and would need to imagine how you would express your verbal position, you might say “I totally agree”. Subsequently, if asked to think about how many stages



you feel you have between feeling ‘neutral’ and feeling that you ‘totally agree’ with an issue, you might feel as though you have two stages between feeling ‘neutral’ and feeling that you ‘totally agree’. In this example ‘neutral’ would correspond to 0 and ‘totally agree’ would correspond to 3. This exercise would have been more complicated and may have yielded response categories that were less personally-meaningful, had you been asked to define the numeric *before* the verbal endpoints. It was deemed prudent, therefore, to first have respondents conceptualise their rating-scale verbally, before having them numerically anchor it.

c. Providing a ‘neutral’ point.

Frequently, rating-scales are bipolar, a neutral point anchored, with the researcher communicating two polar extremes anchored at either end. The neutral point sits equidistant between each extreme, and is often labelled ‘neutral’. From the literature, it is apparent that not all constructs possess a “true” neutral point. However, LTRSs generally include a neutral point – that is a point that allows the respondent to neither agree nor disagree with the statement being evaluated. As such, it was envisaged that the IRSP would have an anchor of origin in the form of a neutral position. This enabled respondents to individualise rating-scales starting from a neutral point, and ensured that a neutral opinion could be expressed about the construct of interest.

Furthermore testing the feasibility of an individualised rating-scale whilst measuring the concept of agreement/disagreement was deemed *useful* given it is a concept so frequently captured by survey research in both academia and industry. It was also more interesting, in terms of gaining insights into respondents’ cognitions, and thus the concept of ‘agreement/disagreement’ was used to aid in the technique’s development.

d. Encouraging equidistant intervals.

Usually, the section of line between each endpoint is divided into intervals, of *numerically equal length* (and often anchored verbally or numerically). Likert-type rating-scales assume that these numerically equal intervals are also conceptually equal. While this assumption does not always hold with LTRSs when all the response categories are labelled (Anglemar and Pras, 1978), it is generally assumed to be a valid assumption when only the endpoints and neutral point are labelled. In addition, as it was thought burdensome for respondents to have to think about whether their different stages on the agreement/disagreement continuum were equidistant, this assumption was maintained within the IRSP. If equidistant intervals had not been assumed it would have substantially increased the complexity of the IRSP, increasing the length of time taken (which is counter-productive) to develop an IRS. It would have also presented a significant problem to researchers when analysing the data. As such, it was necessary to assume that the intervals between the neutral point and each endpoint were equidistant (following the standard Likert-type rating-scale assumption).

e. Mapping the IRSs to allow inter-respondent comparisons.

Given that respondents could be defining IRSs of differing lengths, due consideration needed to be given to how ratings from different IRSs could be compared and analysed. The key point here is that it would not be pragmatic to facilitate individualised rating-scales if the ratings obtained from them cannot be either combined or compared.

The first area to consider was how to map the end-points of the rating-scale. Given that the endpoints of the IRSs are assumed to be conceptually equivalent it would be

possible for the same numerical response to represent different levels of agreement (or disagreement). For example, should respondent A define an IRS of -2 to +2, and respondent B define an IRS of -4 to +4, a rating of -2 would be *conceptually* different across the two respondents (see Figure 4. 1). For respondent A, -2 would represent strong disagreement with an item. For respondent B, -2 would sit halfway *between* ‘neutral’ and strong disagreement. As such, this example demonstrates that it would be conceptually incorrect to say that a rating of -2 from the first IRS is equal to -2 on the second IRS.

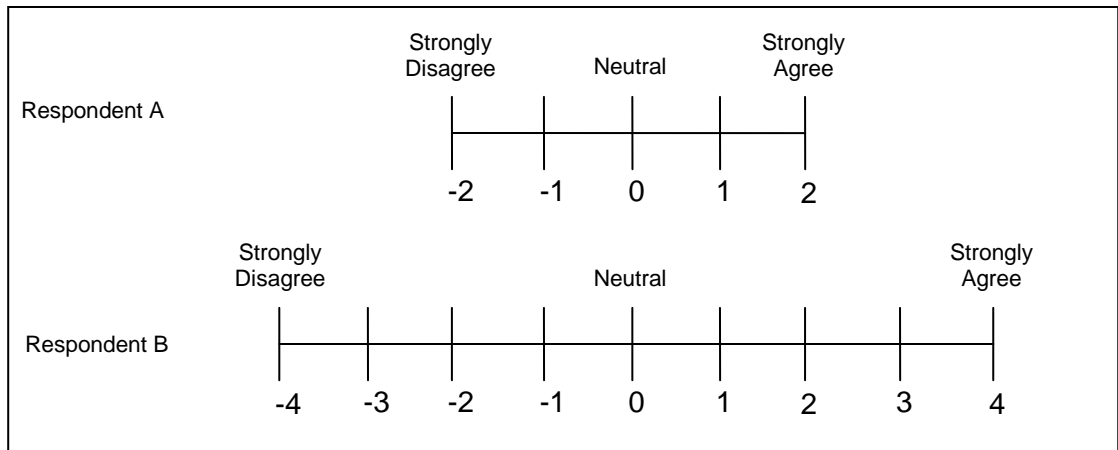
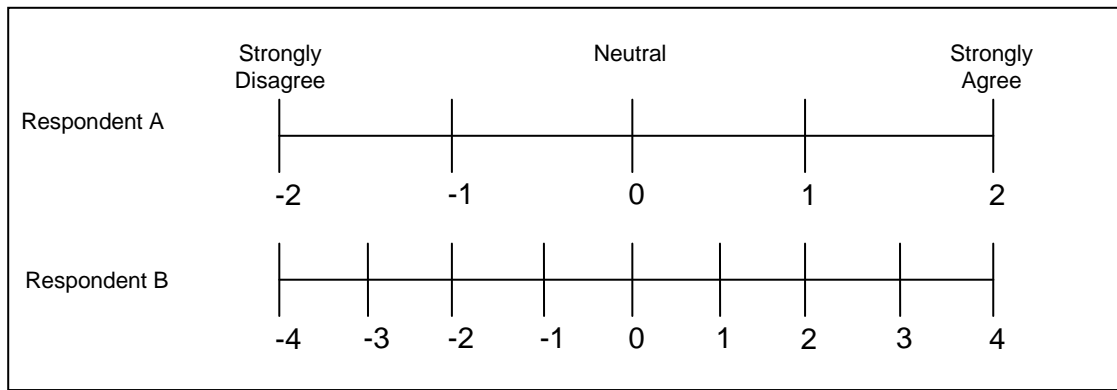


Figure 4. 1 Example using differing IRSs

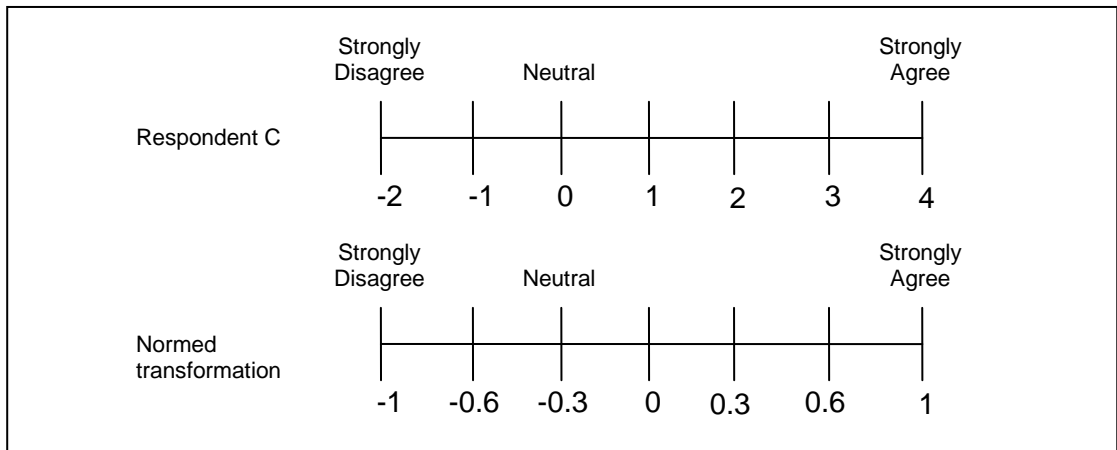
In order to maintain the conceptual equivalence of both IRSs, a rating of -2 from respondent A would correspond to a rating of -4 from respondent B, as shown in Figure 4. 2.



**Figure 4. 2 Example using differing IRSs that are aligned**

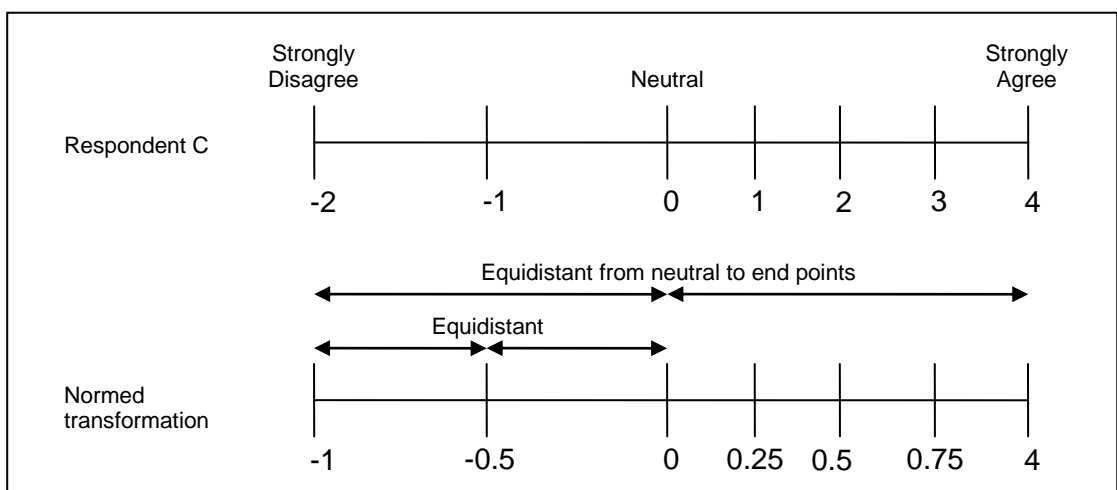
Given the need to maintain conceptual equivalence across IRSs, it was decided that ratings would be transformed into a standardised index from -1 to 1. In the previous example, respondent A's -2 would become -1 when transformed. Whereas respondent B's -2 would become -0.5 when transformed. This would enable the conceptual equivalence of the responses to be maintained across IRSs, thus permitting comparison of ratings and further analysis of all scores.

It was important to consider what would happen in the context of imbalanced IRSs, where, for example, a respondent may desire more intervals for agreement than for disagreement. Should respondent C define an imbalanced IRS of -2 to +4 and if it is regarded as having equal intervals across the whole rating-scale range, then the transformation of the scale would result in a 'neutral point' of -0.3 (see Figure 4. 3). That is, the response on the IRS that was conceptually defined as the neutral point when developing the rating-scale, would suggest mild disagreement on the standardised scale. As each end of the IRS is anchored from the neutral point, this would be a conceptual misrepresentation of the respondent's position on the continuum. In addition, this would be counter to the requirement of encouraging equidistant intervals (abovementioned) where the assumption of equidistant intervals occurred between the neutral point and the endpoints, not between the two endpoints.



**Figure 4. 3 Example using an imbalanced IRS**

In order to conceptually represent the respondent's ratings on the continuum, the interval ratings on the original IRS would need to be treated as two unipolar continuums joined at the neutral point. This means that, for an imbalanced IRS, the distance between neutral and each endpoint are regarded as conceptually equal. In this way, respondent C's imbalanced IRS would have a smaller numeric difference between the intervals on the agreement pole, than on the disagreement pole (as shown in Figure 4. 4).



**Figure 4. 4 Example using an imbalanced IRS transformed appropriately**

In this way, a linear transformation of the response categories, on either side of the neutral point, is possible. The resulting standardised rating-scale is one where the distance between numeric values is equivalent (i.e., the difference between -.2 and -.4 is the same as the distance between .5 and .7) – that is, it fulfils the requirements of equal intervals across its entire range (i.e., -1 to +1). The intervals on the standardised index are conceptually appropriate with regard to the numerical representation of the magnitude of agreement/disagreement felt by the respondent and mapping in this way maintains the conceptual equivalence of the levels of agreement (or disagreement) in relation to each respondent's most extreme level of agreement (or disagreement). It also allows for comparison across respondents to be based on conceptually equivalent values. The conceptual and verbal anchoring stages will need to support this assumption by correctly calibrating respondents in this fashion.

From the above, the following crucial assumptions form the 'starting point' for an individualised rating-scale procedure, namely:

- a. The endpoints of the IRSs are conceptually equivalent as the maximum magnitudes for respondents' agreement/disagreement.
- b. The familiar horizontal line provides a valuable pictorial aid for respondents.
- c. A respondent should verbally anchor endpoints *before* numerically anchoring them.
- d. Respondents' individualise two *bipolar rating-scales* with a neutral point in order to rate their position on the *agreement/disagreement continuum*, rather than a single uni-polar scale from endpoint to endpoint.
- e. 'Adjacent' rating-scale intervals between the neutral position and each absolute endpoint are equidistant.

- f. Using a linear transformation to map scores from each IRS to a normed rating-scale, with a neutral point of zero and endpoints of  $\pm 1$ , will allow direct comparisons within and between respondents.

#### **4.2.2 Development of the ‘rudimentary’ IRSP**

The ‘starting point’, informed the creation of the ‘rudimentary’ exercise which was used in the first round of qualitative data collection. Figure 4. 5 shows the ‘rudimentary’ IRSP instruction sheet, which was used in Round 1 interviews.

##### *4.2.2.1 Visual aid*

A visual aid – a horizontal line – was included as part of the ‘rudimentary’ instruction sheet with a small vertical line at its centre to demarcate the neutral position. The verbal label ‘neutral’ was also added above this marker. Below the marker, a small box was placed with the number ‘0’ positioned inside. This showed the respondent that the neutral position was already anchored numerically at ‘0’. Small vertical lines were positioned to mark the endpoints of the rating-scale horizontal line (as shown in part C of Figure 4. 5).

Instructions

A → Please read the page labelled 'Statements' provided.

B → Where you neither agree, nor disagree with a statement, your position is neutral. In other words, the neutral position is where there is a complete absence of any level of agreement or disagreement with a statement. The number 0 will represent your neutral position.

C →

(c).....disagree                      neutral                      (a).....agree

(d)                       0                      (b)

D → Please think of a word (adverb) that best describes your highest possible level of agreement with a statement. Please write this word on the above line in the space labelled (a).

E → Thinking about the word you've written in (a) above and given that neutral is 0, please assign a number which you feel would best represent the above level of agreement. In other words, if your opinion towards something was statement (a), how many degrees from 0 (neutral opinion) would best represent your view? Please write this number in the box labelled (b).

F → Please think of a word (adverb) that best describes your highest possible level of disagreement with a statement. Please write this word on the above line in the space labelled (c).

G → Thinking about the word you've written in (c), while bearing in mind the number you've assigned to your highest level of agreement and given neutral is 0, please assign a number which you feel would best represent the above level of disagreement. In other words, if your opinion towards something was statement (c), how many degrees from 0 (neutral opinion) would best represent your view? Please write this number in the box labelled (d).

H → You now have your own personally-defined measurement scale. Please use this scale to show how strongly you agree or disagree with the statements you read previously. Please do so by writing the number from your scale which best corresponds with your opinion in the spaces given next to each statement.

**Figure 4. 5 IRSP Instruction Sheet IRSPr1: Interviewees 1 and 2**

There needed to be spaces for respondents to add numbers and verbal labels to the endpoints. Respondents were first asked to picture the conceptual meaning of the endpoints (for them), then add verbal meaning to these (D and F in Figure 4. 5), and lastly to clarify them numerically (E and G in Figure 4. 5). The instruction-order was designed to help respondents to do this. In addition, it was decided that respondents should focus their attention on one 'pole' before proceeding to the other. As such, each respondent would need to verbally and numerically anchor one endpoint, before



considering the other. This was to encourage respondents to numerically anchor each endpoint in a more meaningful way, given that after verbally anchoring the endpoint they immediately attach a numerical clarification to it. A respondent is in a far better position to numerically anchor the endpoint whilst already thinking about it verbally. Next came the question of *which* endpoint, positive or negative, to begin with. It was decided that by starting with the ‘agreement’ pole, thinking about this ‘positive’ emotion first might better ease the respondent into the rest of the process.

Given the above reasoning, two spaces were created at each endpoint, one above and one below the small vertical lines (C in Figure 4. 5). The two spaces above the endpoints were designated for verbal labels. The “.....disagree” was placed on the left pole and the “.....agree” was placed on the right pole, in line with bipolar rating-scales some respondents may already be familiar with. This meant that it would be clear to respondents which side of the neutral point represented the ‘agreement’ pole and *which* represented the ‘disagreement’ pole. These two spaces are where respondents would verbally anchor using an adverb what best describes their absolute level of agreement/disagreement. The two spaces below the rating-scale endpoints were designated for numbers and consisted of two small boxes. In order for there to be some way for the instructions to communicate clearly to the respondent which space to write in, each dotted line and box was labelled. Numbers were deemed unsuitable given the potential to confuse or influence the respondent when it came to them numerically anchoring the endpoints. As such letters ‘(a)’, ‘(b)’, ‘(c)’ and ‘(d)’ were chosen.

#### 4.2.2.2 *Instruction order*

To aid respondents they were first presented with a short list of statements to read before individualising their rating-scales (A in Figure 4. 5). This was done to help them generate rating-scales that were more meaningful to them. This ‘warmed up’ the respondent’s mind to think about *how* they agree/disagree and was designed to help respondents think about the number of agreement/disagreement categories they have. Secondly it was important that respondents be made to understand the concept of ‘neutral’ *before* being presented with subsequent instructions (B in Figure 4. 5). However, it was useful that the visual aid should *follow* the definition of ‘neutral’ so that it would start to bring the visual aid ‘to life’ with regard to the continuum it represented (C in Figure 4. 5). In addition, there needed to be an instruction to help respondents verbally anchor the ‘agreement’ endpoint in the space labelled ‘(a)’. This had to include a prompt to help respondents visualise their absolute agreement and an instruction to attach a verbal meaning to this (D in Figure 4. 5). Moreover, it was decided that respondents would need to be prompted to numerically anchor the ‘agreement endpoint’ by writing a number in the box labelled ‘(b)’ (E in Figure 4. 5). The same process then needed to be repeated for the ‘disagreement endpoint’ (F and G in Figure 4. 5). Finally, respondents were instructed to use the response categories they had created, to rate their opinions on the statements previously presented (H in Figure 4. 5). This completed the IRSP instruction sheet for use in Round 1 interviews.

The instructions in Figure 4. 5 take the respondent through eight steps to create their rating-scale. These instructions formed part of the ‘rudimentary’ IRSP, which needed to be tested for its effect on respondent choices. For a more detailed explanation of why specific instruction wording was used please refer to Appendix A. In addition, a set of

items were included for the respondents to rate, enabling them to experiment with their individualised rating-scales.

### **4.2.3 Interviews: Data Collection and Analysis**

The approach taken towards the data collection and analysis is outlined in the Methodology chapter.

The commitments of constant comparison and theoretical sampling, albeit somewhat purposeful in this case, meant that the entire qualitative process was highly interactive and iterative, with an absence of the traditional distinction between data collection and data analysis. Data analysis proceeded as soon as sufficient material was collected to work on (rather than waiting until a predefined data set was obtained), and this in turn fed back into the collection of new data. This dynamic relationship between data analysis and data collection was a critical characteristic of the whole approach.

In doing this research, the externalisation of the data analysis and reflexivity required by grounded theory, meant that data were always recorded in several ways: specifically memos, notes on interview protocol sheets, and node descriptions in NUD\*IST N6 were used. This made transparent the full interpretive processes of knowledge production. This chapter brings together this material, especially concerning the decisions made for individual modifications to the IRSP. Not only are the findings that led to the IRSP modifications outlined, but so are the insights gained about interviewees' conceptualisations and interesting patterns observed (regardless of whether they led to any modifications). In this way, an account of what the researcher noticed and interpreted is expressed.

#### **4.2.4 Interview Protocol**

The research objective<sup>1</sup> was used to inform an interview protocol covering the key areas to be explored. The points in the interview schema are outlined in Figure 4. 6.

---

<sup>1</sup> To test the feasibility of whether respondents can self-define a rating-scale using the IRSP, such that (a) they completely understand the nature of the task, (b) they can define and use it with ease, and (c) all the intervals on the rating-scale are personally meaningful.

- A. Seek to probe as deeply as possible into how instruction-wording influenced their respondent choices.
  - a. Why did they choose their verbal endpoints?
    - i. If the respondent chose different verbal anchors (e.g. 'totally agree' and 'completely disagree'), explore why.
  - b. Why did they choose their numerical endpoints?
    - i. If the respondent chose a numerically imbalanced rating-scale (e.g. two intervals on agreement, and three intervals on disagreement) explore why.
- B. Did the respondent encounter any problems/ambiguities with the chosen practice items?
- C. Does the respondent feel as though the rating-scale accurately reflects their views?
  - a. Are each of the intervals meaningful to them?
    - i. If not, explore why.
    - ii. How might the instructions be modified so that they would lead others into defining more meaningful rating-scales?
    - iii. How would they change their rating-scale to make it more meaningful?
  - b. Do the endpoints represent their entire span of agreement/disagreement?
- D. How does the respondent feel about the respondent-defined rating-scale process?
  - a. Clarity of instructions?
  - b. Ambiguities?

**Figure 4. 6 Interview Protocol Sheet**

A reminder was also included in the schema with regards to the timing of certain questions, given that bias could have been introduced into some of the interviewees' responses depending on the point at which each question was asked (see Table 4. 1).

**Table 4. 1 Planned order of activity and probes for Interviews**

Task	Time (progresses from left to right)						
Interviewee is asked to read a list of statements.							
Interviewee reads and executes IRSP instructions.							
Explore probes in bullet point A.							
Interviewee is asked to use the rating-scale they have just designed to rate the previously read statements.							
Explore probes in bullet point B.							
Explore probes in bullet point C.							
Explore probes in bullet point D.							

For example, asking an interviewee questions about the meaningfulness of their chosen intervals *before* they have used it to rate the statements, would render them more acutely conscious of their rating-scale and might alter the natural way they would use it to rate items. This is why consideration was given both to what questions to ask and when to ask them.

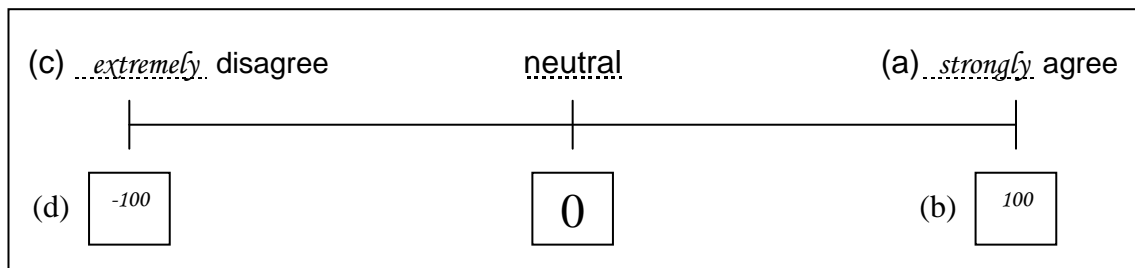
**4.2.5 Round 1 – Interviews 1 and 2**

The instruction sheet shown in Figure 4. 5, was used in Round 1 interviews (hereafter referred to as IRSPr1), of which there were two. The process was so preliminary in nature that the first two interviews provided enough insight for improvements to be made to the initial instructions.

*4.2.5.1 Key findings*

*Individualised Rating-Scales (IRSs) chosen*

Interviewees 1 and 2 defined the following rating-scales:



**Figure 4. 7 Interviewee 1: IRS Chosen**

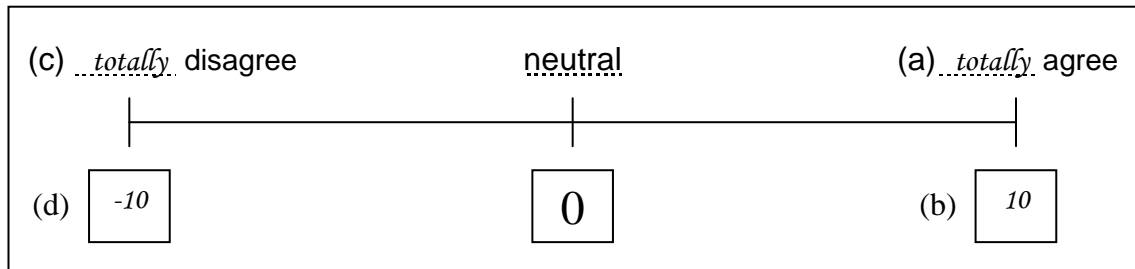


Figure 4. 8 Interviewee 2: IRS Chosen

*Distinctiveness of response intervals*

Not all the response categories on the Individualised Rating-Scales (IRSs) defined by both interviewees were *meaningful*. Interviewee 2 chose an IRS with parameters  $-10 \leftarrow 0 \rightarrow 10$ . When prompted as to *why* he chose these numbers he found it difficult to explain. However, he suggested he may have chosen that particular number because he thinks about many things being rated ‘out of 10’; “girls” and “cars”. Whilst the examples he gave are humorous, it makes for some interesting insight. He essentially saw the number ‘10’ as representing a ‘maximum level’, and he may have subconsciously applied this logic when defining his IRS. However, when asked whether he could meaningfully differentiate between a response of 7 and a response of 8, it proved difficult for him. Interviewee 1 had an IRS with the parameters  $-100 \leftarrow 0 \rightarrow 100$ . During the interview she was prompted to discuss why she had chosen these numerical endpoints, to which she responded by explaining that the number ‘100’ meant that she felt as though she was ‘100%’ agreeing/disagreeing with something. The discussion indicated that rather than being prompted to count the number of cognitive stages between feeling ‘neutral’ and feeling that she ‘strongly agreed’ with something, she jumped straight into thinking about the concept of ‘strongly agreeing’ and equated it with ‘*feeling 100%*’ about something; resulting in the selection of ‘100’ as her

numerical endpoint. She indicated that she chose ‘-100’ because she simply applied the same thinking to the other side of the ‘neutral’ position (i.e. the disagree pole), thinking of it as a “mirror-opposite”.

These findings indicated that the instructions for the numerical anchoring needed to be improved further, so that respondents would be encouraged to produce more distinct response intervals. This would mean that the instructions would need to guide them into thinking about how they anchor their endpoints in a different way. However, it was clear that more data would need to be collected (with the numeric-anchoring-instructions as they are) before this modification could be made. Modifications would be put on hold until further light could be shed on this issue. A note was made to probe this area further in subsequent interviews.

#### 4.2.5.2 *Key modifications*

##### *Verbal anchoring instructions*

Both interviewees were able to choose an adverb to place before ‘agree’/‘disagree’. There was a small degree of difficulty in that Interviewee 1 asked whether the aim was to “put a word that meant “agree”” (i.e. a synonym for ‘agree’) in the space provided. She was instructed to execute the instruction however she thought best, and that there was no ‘right’ or ‘wrong’ way of doing the exercise, because the point was to observe someone’s *interpretation* of the instructions. She re-read the instructions before deciding against her initial interpretation, and wrote the word ‘strongly’ next to ‘agree’. When asked whether she found any part of the instructions confusing or ambiguous, she responded by saying that she found some of the wording over-complicated and that some of the sentences were a bit “long-winded”. This meant she had to re-read several



instructions. When asked to highlight these particular areas she indicated the section of text that read “the highest possible level of agreement”. She explained that she took it to mean “the most I could agree with something”, after re-reading it.

This finding indicated that the instructions for the verbal anchoring needed to be improved.

Modification: The phrase “your highest possible level of agreement” (which appears in D and F in Figure 4. 5) was replaced with “the most you could possibly agree”. This addressed the above issue raised with regard to the original phrasing. Interviewee 1’s abovementioned explanation of how she interpreted the original phrase was useful, in that, the new phrase was grounded in her explanation of what she thought the instruction had meant. It was an improvement because using the word ‘highest’ in the original phrasing had the potential to subconsciously bias a respondent into choosing the adverb ‘highly’ (whereas it was deemed somewhat less likely that someone would choose the word ‘mostly’ to represent their *absolute* stance).

Interviewee 1, whilst she chose ‘100’ and ‘-100’ as her numeric endpoints, she chose different verbal endpoints. She was probed on this issue. Interestingly she indicated that she thought that she was not permitted to choose the same verbal label twice. It is worth noting that there is no part in the IRSPr1 that indicates one must choose a different adverb – in fact this point was not raised at all in the instructions. Yet, she still assumed this to be the case. She was asked, whether, if the instructions had included a sentence that led her to believe that she could have chosen the same word again, whether she

would have done. She indicated that she would have used the same adverb on both verbal endpoints. Even if only a small minority of respondents could potentially interpret the instructions in this way, it was deemed prudent to control for this happening at all.

These findings indicated that a sentence needed to be added to remove any ambiguity as to whether the same adverb could be chosen again.

Modification: “This word can be the same as or different to the word you wrote in (a)” was added (to F in Figure 4. 5).

#### *Inclusion of Greenleaf’s sixteen items in the IRSP*

It became clear that observing the interviewees’ use of the IRS to rate a scale of items, provided a picture of their opinion about that construct, but offered limited insight into the meaningfulness of each of the intervals of their IRS. In other words, an interviewee might feel positive about a construct under measurement and would therefore mostly agree with the items presented to them, using mainly one pole of their IRS to rate the items. This meant that interviewees could only be observed using a few of the intervals available to them on their IRS. Whilst some knowledge on how they are using their IRS could be gained, it would have been more useful to observe the interviewees using the full range of their rating-scale, so as to better determine the meaningfulness of the intervals. In order to do this it was necessary to incorporate a list of items that did not correspond to a particular construct or latent variable. In other words, it was necessary to have as wide a variety of likely responses as possible.

Modification: Incorporated Greenleaf’s (1992b) bank of sixteen uncorrelated items.

Greenleaf uses this set of uncorrelated items to measure response styles.

These items are highly uncorrelated. In other words, a respondent would be expected to experience varying levels of ‘agreeing’ and ‘disagreeing’ across these sixteen items. It was decided, therefore, that these sixteen items provided the ideal ground for which to observe interviewees using the full range of their IRSs. As such, Greenleaf’s list of items was incorporated into the exercise for Round 2.

The modified instruction sheet for the IRSP can be seen in Figure 4. 9 (hereafter referred to as IRSPr2), and includes the abovementioned modifications. Respondents were provided with Greenleaf’s sixteen items to rate in round two (Interviews 3-7).

#### **4.2.6 Round 2 – Interviews 3-7**

Round 2 consisted of five interviews (3, 4, 5, 6 and 7). Each interview was tape-recorded (permission having been granted in all cases) using a Sanyo Cassette Recorder. Each tape was transcribed and coded using NUD\*IST N6<sup>2</sup>. Please refer to Appendix B for one of the interview transcripts (the other transcripts are available on the CD accompanying this thesis), and Appendices C and D for a breakdown of the tree nodes created to code the data in NUD\*IST N6.

---

<sup>2</sup> Software for qualitative data analysis.

Instructions

A → Please read the page labelled 'Statements' provided.

B → Where you neither agree, nor disagree with a statement, your position is neutral. In other words, the neutral position is where there is a complete absence of any level of agreement or disagreement with a statement. The number 0 will represent your neutral position. This has been labelled on the line below.

C →

(c).....disagree                      neutral                      (a).....agree

(d)                                             (b)

D → Please think of a word (adjective) that best describes the most you could possibly agree with a statement. Please write this word on the above line in the space labelled (a).

E → If your opinion towards something was statement (a), think about how many steps away from 0 (the neutral opinion) would best represent your view? In other words, please assign a number which you feel would best represent the level of agreement you have described in statement (a). Please write this number in the box labelled (b).

F → Please think of a word (adjective) that best describes the most you could possibly disagree with a statement. This word can be the same as or different to the word you wrote in (a). Please write this word on the above line in the space labelled (c).

G → If your opinion towards something was statement (c), think about how many steps away from 0 (the neutral opinion) would best represent your view? In other words, please assign a number which you feel would best represent the level of disagreement you have described in statement (c). Bear in mind that this number must be negative. Please write this number in the box labelled (d).

H → You now have your own personally-defined measurement scale. Please use this scale to show how strongly you agree or disagree with the subsequent statements. Please do so by writing the number from your scale which best corresponds with your opinion in the spaces given next to each statement.

**Figure 4. 9 IRSP Instruction Sheet IRSPr2: Interviewees 3-7**

*4.2.6.1 Key findings*

*Individualised Rating-Scales (IRSs) chosen*

The IRSs chosen by interviewees 3-7 have been summarised in Table 4. 2. Shown, are the verbal labels and numerical anchors chosen.

**Table 4. 2 Interviewees’ 3-7: Numerical and Verbal Endpoints Chosen**

Interviewee	Verbal		Numeric		Changes Numeric	
	+	-	+	-	+	-
3	Completely	Totally	10	-10	5	-5
4	Emphatically	Absolutely	10	-8	none	none
5	Totally	Totally	10	-10	1	-1
6	Completely	Completely	10	-10	none	none
7	Strongly	Strongly	4	-4	none	none

The columns labelled ‘Changes Numeric’ show whether the respondents would have changed their numerical anchors after rating Greenleaf’s sixteen items using their IRS. ‘None’ means that the respondent was happy with their original IRS and did not feel the need to change their numerical anchors. A number in this column indicates the desired change to the anchor. For example, in retrospect Interviewee 3 would have preferred to have numerically anchored her scale from -5 to 5:

- \*Respondent 3: But after doing the questionnaire I now wish that I’d chosen five –
- \*Interviewer: Ok –
- \*Respondent 3: As that would’ve helped me hone down...
- \*Interviewer: That’s interesting.
- \*Respondent 3: Because conceptually the one to ten, it has kind of too many little increments...
- \*Interviewer: Right that’s good, yes...
- \*Respondent 3: And you need to think ooh is that an eight or a nine, I don’t know –
- \*Interviewer: Yes.
- \*Respondent 3: But if it’s the difference between a 3 and a 4 [on a -5 to 5 scale] it’s kind of like a big chunk.
- \*Interviewer: Yep.
- \*Respondent 3: So you can kind of get to grips with that I think more.

[Interview 3: 260-284]

It is quite clear from this extract that she did not find all the intervals on her -10←0→10 IRS distinct, and her use of the word ‘chunk’ expressed that one could look at it in terms of ‘chunks’ being ‘cut too finely’ when she had too many increments. It is worth mentioning that none of the interviewees expressed a desire to change their verbal

anchors when prompted, and none wanted to *increase* the number of intervals in their IRS.

#### *IRSP execution times*

Table 4. 3 shows the time it took each interviewee to complete the process of creating their own IRS. The times were measured by listening to the taped interviews and recording the time taken between each stage. A stop watch was not used as this may have influenced the respondents. The exercise took an average of 5 minutes and 23 seconds. This was measured to determine how much extra time would be added to expected survey completion times, if respondents were using the IRSP to report their answers. As this point in time, it was *not* considered to be taking respondents a long time to execute.

**Table 4. 3 Interviewees' 3-7 Exercise Completion Times**

Interviewee	Greenleaf statements		IRSP (Reading & Executing)	Total (mins)
	Reading time (secs)	Rating time (secs)	Time (secs)	
3	35	120	160	5.25
4	35	90	150	4.58
5	45	130	120	4.92
6	35	160	120	5.25
7	34	135	200	6.15
<b>Avg time (secs)</b>	<b>36.8</b>	<b>127</b>	<b>150</b>	<b>5.23</b>

#### *The mystery attraction to '±10'*

As can be seen from Table 4. 2, four out of the five interviewees were attracted to '±10' for either the positive or negative numeric endpoints, or *both*. This appeared to echo Interviewee 2's attraction to '±10'. On probing this choice, interviewees gave some of the following explanations:

[Interview 3: 272-273]

“Umm...I don't know...it's just a kind of a western counting – there is a word – it's just a standard digital type of thing really.”

[Interview 5: 180, 184, 315-316]

“I do things in evens, and 10 is always the scale, like 10+ and 10-.

[...]

It’s a nice even big number.

[...]

Because it’s 0 to 10, that kind of gives you your 10% agree with it, your 20% agree with it, so 10 would be 100%...that’s kind of why I chose it –”

As shown in Figure 4. 10, Interviewee 5 would have changed his numerical anchors (from  $\pm 10$  to  $\pm 1$ ) so that his IRS was more meaningful. The graph in Figure 4. 10 shows the spread of his responses to the sixteen items using his IRS.

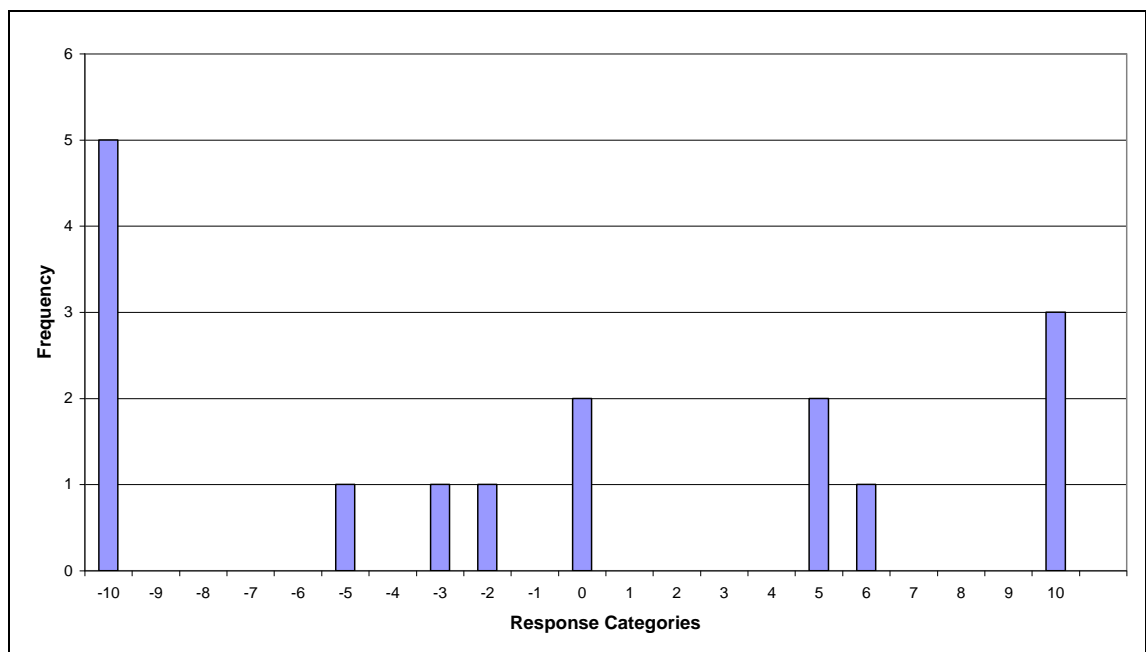


Figure 4. 10 Interviewee 5 – Spread of responses

As can be seen, approximately two thirds of his responses were either 10, 0, or -10. He failed to use fourteen out of the twenty-one intervals on his IRS. It was quite clear that this individual’s IRS had more intervals than were needed. Interviewee 5 also expressed that he did not genuinely feel as though he needed any points between neutral and either extreme. When further questioned about this, it was brought to his attention that there were items which he rated using numbers *between* 0 and  $\pm 10$ , and that a rating-scale of  $\pm 1$  would not have permitted him to choose a ‘milder’ form of (dis)agreement for those

statements. He indicated that he had only rated those statements in this way because “they were hard questions” but in retrospect he feels that he totally agrees/disagrees with each of them and that a -1 to 1 rating-scale would have served his needs well. When asked about how the instructions could have been worded differently (so that people would choose more meaningful rating-scales), he said;

- \*Respondent 5; [...] well you ask if you were to agree with something to your maximum agreement, write down a word – ask them to right down a word, which would be, if you were sort of, were unsure about something –
- \*Interviewer; Right –
- \*Respondent 5; Or halfway agreed, what word would you put to that.
- \*Interviewer; Ok –
- \*Respondent 5; And then maybe they could assign a number to that –
- \*Interviewer; Yeh –
- \*Respondent 5; So then that would make sense in the scale, so they’d have neutral, midway and then total agreement and you’d have numbers for each. So you could probably get a better scale –

[Interview 5: 327-346]

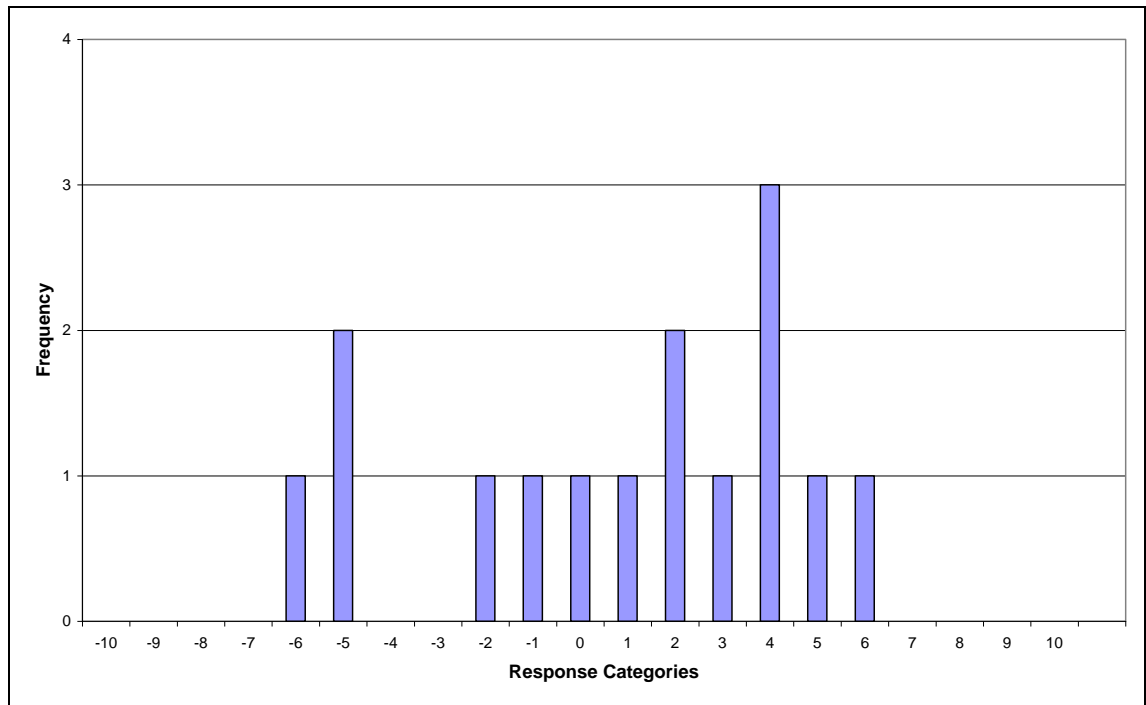
Interviewee 5, was suggesting that respondents be prompted to think about ‘midpoints’ (referring to points *between* neutral and either endpoint), in much the same way as he was made to think about his endpoints. Whilst this suggestion was considered, it was decided that it was likely to be somewhat problematic for two main reasons. Firstly, the length of time required to develop the IRSP would increase significantly and the time it would take to read and execute it would increase concomitantly. Should it become a process that takes too long, this would make it redundant as a useful measurement instrument. Secondly, prompting respondents to conceptually think about a ‘midpoint’, and assign a verbal meaning to it, might lead them into defining a rating-scale with only one response category between neutral and each endpoint. This is likely to result in many respondents choosing an IRS of  $\pm 2$ , which would not be ideal if several of them could have used meaningful rating-scales with more intervals (and therefore greater



information-transmitting capacities). As such, Interviewee 5's suggestion was not executed literally.

However, the suggestion *did* shed light on a potentially superior way of having respondents individualise their rating-scale lengths. This interview script prompted the idea of respondents numerically anchoring their IRSs in a slightly different way. It involved the notion that focusing on the number of steps *in-between* neutral and both endpoints would be clearer to respondents. In other words, it was considered whether respondents could assign a number representing the *number of steps in-between* neutral and their two extreme positions, as opposed to assigning a number which represents the *extreme position* (as was currently being done). This idea was noted, and eventually led to the creation of a second version of the IRSP (IRSPv2). This is further discussed at the point it was developed.

Interviewee 6 chose a scale from 10 to -10. Interestingly though she did not use any positive or negative 7's, 8's, 9's or 10's when rating the statements, as can be seen from the graph in Figure 4. 11.



**Figure 4. 11 Interviewee 6 – Spread of responses**

When asked if she would change her rating-scale if given the option she replied by saying "You want them to be lower?" [Interview 6: 89]. This was not prompted, in that she was not told *how* she could change the rating-scale. Clearly, therefore, shortening it may have been something that had crossed her mind. She was instructed that there was no expectation either way, in order to assure her there was no 'right' or 'wrong' way to respond to the question. She then stated that she probably would not change her rating-scale at all, adding that,

"\*Respondent 6: Yeh, I don't think I feel particularly, really strongly about any of the statements."

[Interview 6: 159-161]

This was her way of justifying why she did not use any response categories above  $\pm 6$ . Whilst her reasoning may be true, it was considered possible that she might possess some form of response style (e.g. *mid-point responding*), which would explain why she chose not to show extreme opinion. However, later she stated that had the items been

addressing themes such as ‘murder’ or ‘death’, she would have had very strong opinions. Whilst this might in fact be an accurate observation on her part, it was considered that perhaps the IRS chosen was marginally longer than her ideal rating-scale (i.e. present in her mind), as she mentioned that her reason for choosing an IRS of  $\pm 10$  was,

"I don't know, most scales that you think of numerically are in the thing of 1 to 10 - or whatever - so I just thought I could use that [laughs]."

[Interview 6: 112-113]

This highlights a potential external influence that prompts respondents to choose  $\pm 10$ . However, the reason her observation was considered to have some credence is because her responses are fairly uniform across  $\pm 6$  (on items for which she claimed did not cause her to feel particularly strongly). In addition, if a respondent is genuinely capable of feeling meaningful discrimination between each interval on a twenty-one point scale, they would of course be unable to use every one of their intervals when rating only sixteen items. So on a rating-scale of  $\pm 10$ , there would always be at least five intervals left unused when examining responses to Greenleaf's sixteen items. The question is more about whether clusters are clearly forming, as was done by Interviewee 5 who had 50% of his responses clustered into 10 and -10. The same cannot be said for Interviewee 6.

Interviewee 4 chose a rating-scale of -8 to 10, and stated that all the categories were meaningful to him. He argued that some of the intervals were unused because his opinion to the statements were not varied enough to warrant choosing those particular intervals;

"Umm...I suppose the scale could be too long but, however, I don't feel as though some of the questions actually required a big agreement or disagreement, or a very small one...either way...there's no real medium in-between the majority of questions."

[Interview 4: 204-207]

However he did not use any 7's, 2's, 1's and 9's when rating the statements. So despite his reasoning - that the statements he rated did not warrant those positions - the rating-scale appeared to be too long for him. He did acknowledge this was possible.

On the whole, it seemed that respondents who chose endpoints of  $\pm 10$  were doing so out of an external predisposition, rather than as a result of meaningful consideration for each interval from neutral to their endpoints. Therefore it was assumed that IRSs of  $\pm 10$  might be an indication that respondents are not fully engaging in the directives, and therefore produce rating-scales lacking personally-distinct intervals.

#### 4.2.6.2 *Key modifications*

##### *Greenleaf's Items*

It was clear that several of Greenleaf's items would need to be modified for round three of the qualitative interviews. There were ambiguities and misinterpretations of certain words contained in the items. For example, almost all the interviewees had a problem understanding what was meant by the term 'homebody' in item two, "I am a homebody."

“\*Respondent 4; Umm... ‘homebody’, what’s that?”

[Interview 4: 27-28]

“\*Respondent 5; Uh...I don’ really understand that [he points at the word ‘homebody’ in statement 2]”

[Interview 5: 18-20]

“\*Interviewer; Yes. Was there anything you didn’t understand?  
 \*Respondent 6; ‘Homebody’...likes being at home?”  
 [Interview 6: 20-23]

“\*Respondent 7; Homebody?”  
 [Interview 7: 31-32]

Definitions of the term ‘homebody’ were examined, with a suitable one being found on WordNet<sup>3</sup>: The *noun* ‘homebody’ is defined as “a person who seldom goes anywhere, one not given to wandering or travel. Synonym: stay-at-home.”

Modification: The original item was reworded by replacing the word ‘homebody’ and using the above definition to rephrase the item. This helped to avoid further ambiguities in the text. The following phrase became the revised second item: “I generally prefer to stay at home than go out.”

Another area of slight confusion was highlighted in Interview 5:

“\*Respondent 5; That sort of, doesn’t really make sense, but –  
 \*Interviewer; Which one? Question 4? [Respondent is referring to statement 4]  
 \*Respondent 5; Yeh, ‘how fast our income goes up, we never seem to get ahead’...umm...the more money you get?”

[Interview 5: 35-41]

Here the respondent is referring to Greenleaf’s fourth item “No matter how fast our income goes up, we never seem to get ahead.” It was clear that this statement was not suitable for students, as it implies that the respondent is part of a larger family unit at home, with its references to ‘our income’ (i.e. the earnings of ones’ family such as ‘my wife and I’). For this reason it was appropriate to replace the word ‘our’ with the word ‘my’. However, there was still a problem with the contextual relevance with the modified phrase ‘No matter how fast my income goes up...’. Generally speaking,

<sup>3</sup> <http://www.answers.com/main/ntquery?dsid=502&deid=1424192077> accessed on 28<sup>th</sup> May 2006.

students are either on little or no income, and so it is redundant to ask whether students feel as though they ‘never get ahead’ no matter ‘how fast their income goes up’, because it is safe to assume that a student’s income certainly is not ‘going up’. At this point in their life they are not usually trying to ‘get ahead’ financially-speaking. In fact students are usually getting into more debt. The original term ‘goes up’ also confused the respondent quoted above.

Modification: Given this contextual issue, the item was modified to “I never seem to have much money.”

Whilst examining the remainder of Greenleaf’s items it became obvious that some remained very ‘American’ in their origin. It was deemed best to replace these Americanisms with a British equivalent to avoid future ambiguities in the items.

Modification: In this manner, the phrase ‘TV commercials’ in item nine was replaced with ‘TV advertisements’. On consideration of item ten, “A college education is very important for success in today’s world,” the word ‘college’ was replaced with the British equivalent, ‘university’, given that British people usually use the word ‘college’ when referring to sixth-form students and not university students.

### *IRSP Instructions*

Interviews 3-7 highlighted several areas where the second version of the IRSP template could be further developed. Analysing the interviews with respect to problems with the clarity of the instructions, or how they might be improved, revealed several possibilities.

An extract from Interview 3 illustrates this:

“\*Respondent 3; Ummm. Whether I ‘agree nor disagree with a statement my position is neutral’, it was just kind of – I don’t know – a tiny bit confusing.”

[Interview 3: 53-55]

Respondent 3 felt that the first sentence of the instruction sheet was "a tiny bit confusing". She is referring to the sentence in the text that reads "where you neither agree, nor disagree with a statement, your position is neutral". From observing her body language and the way she stressed certain words, it appeared to be a problem with the way this sentence was presented as a double negative (i.e. using words like ‘neither’ and ‘nor’). It seemed that negatively worded phrases made respondents think doubly hard about what the sentence was telling them. For this reason, it was thought best to develop this instruction by changing this sentence or removing it and replacing it with a different description of the neutral position.

On asking Interviewee 3 how she would have phrased it differently (after she had understood what was meant by the instruction), she responded;

“\*Respondent 3; Uh, I think I’d put ‘when answering these questions you start from a neutral position’...ummm...and then with each question you move from 'a' – to 'a', either side of the neutral position...suppose in agreement or disagreement.”

[Interview 3: 65-69]

Interviewee 3 seemed to think that one is in a state of ‘neutrality’, and is then catalysed to move in either one direction or the other (agreement or disagreement) along a bipolar continuum. This might be why the abovementioned instruction confused her, as it talks about an *absence* of agreement/disagreement. Conceptually speaking, the instruction was not in line with the way in which she thinks about ‘neutrality’ and ‘agreement/disagreement’.

In IRSPr2, the ‘neutral’ instruction read as follows (B in Figure 4. 9):

Where you neither agree, nor disagree with a statement, your position is neutral. In other words, the neutral position is where there is a complete absence of any level of agreement or disagreement with a statement. The number 0 will represent your neutral position. This has been labelled on the line below.

This first part of this instruction (underlined) was thought of as the part where the respondent reads a ‘definition of neutral’. The second part of this instruction (not underlined) was thought of as the part where the respondent is shown that ‘neutral’ represents ‘position 0 on the line’. It was considered that the instruction may have been clearer had the respondent been presented *first* with ‘position 0 on the line’, and *then* presented with the ‘definition of neutral’. In this way, a respondent would be reading a definition of the term ‘neutral’ having already seen the visual aid to support their understanding of this definition. Modifying the instruction in this manner would improve a respondent’s immediate understanding of the concept of ‘neutral’.

On consideration of these points, the instruction would be improved, (a) if it were made shorter, (b) if ‘neutral’ was defined in a less complicated way, and by removing the terms ‘neither’ and ‘nor’ from the definition, (c) if a reference was made to ‘position 0 on the line’ before providing a ‘definition of neutral’.

Modification: The instruction was modified to account for these three points raised. As

a result, the original instruction was replaced by;

The number 0 on the scale will represent your neutral position (i.e. no opinion). This has been labelled on the line. This means that if you didn’t agree or disagree with a statement it would be rated 0.



After Interviewee 3 had defined her rating-scale and had read the part that says (H in Figure 4. 9),

You now have your own personally-defined measurement scale. Please use this scale to show how strongly you agree or disagree with the subsequent statements. Please do so by writing the number from your scale which best corresponds with your opinion in the spaces given next to each statement.

She exclaimed, “Oh I’ve made it really hard for myself” [Interview 3: 42]. It was evident that after reading what the purpose of the exercise was, she regretted her chosen endpoints of ‘+10’ and ‘-10’. Her reaction was almost immediate, in that she realised instantly the difficulty of what lay ahead (in that her rating-scale had far too many intervals). Later on in the interview, Interviewee 3 referred back to this point and said that she would rather she had been told at the beginning of the instruction sheet that she was about to personally-define a rating-scale.

“\*Respondent 3; Umm... if it had said something about defining my own measurement scale up here [indicates very top of page]”

[Interview 3; 349-351]

She stressed that the entire exercise “makes complete sense” [Interview 3; 362] but that one would define their scale ‘better’ if they knew a little about the reason behind the exercise. It was suggested to her (purely as a means of playing ‘devil’s advocate’) that perhaps others may not understand what ‘defining a measurement scale’ means, like she does, given she had a past interest in psychology. However, she was adamant it would have improved the exercise, stating that she was sure she would have chosen a much shorter rating-scale had she known what the ultimate purpose of the exercise was.

“\*Interviewer; Right so you feel that if that would have been told in the beginning –

\*Respondent 3; Yeh, I would’ve used 5 [as an endpoint].

...  
 \*Respondent 3; And then I would've known what I was doing kind of thing, but it's a bit *straight in* [referring to present method used].”

[Interview 3; 363-390]

On consideration of this point, the advantages and disadvantages of informing the respondents that they need to ‘define their own rating-scale’, before *doing so*, was considered. Two scenarios were weighed up: Scenario (a) keeping the exercise the same, and Scenario (b) adding a short paragraph explaining that they will be using their instrument to rate the previously read statements to reflect their varying levels of opinion on the agreement continuum.

#### Scenario (a) – No change

The benefit of not telling the respondent the purpose of the exercise could mean that they are potentially less influenced by a predisposition to choose endpoints they have used/seen before, and are therefore familiar with (i.e. typical Likert rating-scales). In other words they might think ‘oh I see’ and access memories of times where they have answered questionnaires and subconsciously/consciously plot those same endpoints. The drawback, is that anchoring their rating-scales, without knowing the purpose of the exercise, is likely to result in less practical rating-scales (as what happened with Interviewee 3).

#### Scenario (b) – Inform respondent of the purpose of the instrument

The benefit here is that the respondent would anchor their rating-scales with more consideration for the purpose of the rating-scale. The drawback is that they would be more likely to access memories of rating scales they have seen in questionnaires and may not focus only on what the instructions are asking of them.

Modification: On consideration of the issues, it was thought best to further develop the IRSP by adding a short paragraph explaining the link between the statements/items read and the instrument they were about to create. For this reason the following instructions were added to the start of the exercise:

Now think about the statements you have just read.

- Think about whether or not you agree or disagree with each statement.
- Also think about how much you agree or disagree with each statement.
- If you have no opinion on a statement then your position is neutral. This means that you don't agree with it, and you don't disagree with it.

It was thought that this would hopefully get the respondents thinking more about the purpose of a rating-scale when they considered anchoring their endpoints. The first bullet point helps them to determine 'which side of the fence they sit on'. The second bullet point helps them to think about 'how far away from the fence they are'. The third bullet point helps them to conceptualise 'the fence' (i.e. 'neutral').

Modification: Subsequent to the above instruction, the title 'define your own opinion scale' was added as a result of feedback received from some of the respondents.

For example Interviewee 3 said "I would've used titles like [...] 'Choose your own measurement scale'" [Interview 3: 381-385]. It was clear the interviewees felt that the addition of a clear title would have shed light on the purpose of the exercise and brought focus to the process. However, accessible words needed to be chosen for the title. It was

appropriate to refer to the instrument as an ‘opinion scale’, because a term such as ‘measurement’ or ‘rating’ -scale was considered to be somewhat too scientific and would be harder for respondents to relate to. Interviewee 4 defined different intervals on a rating-scale as “differences of opinion” [Interview 4: 228]; from the context, this meant ‘different levels of opinion’. The word ‘opinion’ was also mentioned in several of the other interviews (e.g. Interviews 3, 4, 5) when referring to the rating-scales, and so this term clearly seemed readily accessible to respondents. As such, the subtitle ‘Defining your own opinion scale’ was included in bold. This also assured that if any respondents chose to ‘skim-read’ the instructions (which cannot be avoided), they would be likely to notice the purpose of the exercise.

Finally, it was necessary to put the rating-scale visual aid on a separate sheet of paper, so that it could be made larger and clearer to the respondents. The visual aid used in IRSPr2 (C in Figure 4. 9) was part of the instruction sheet. From observing the interviewees, the visual aid appeared too ‘busy’. The verbal anchors were positioned over half their side of the continuum to allow enough space to write in their verbal anchors. Visually, the verbal anchors represent the respondent’s true endpoints, and the visual aid in IRSPr2 did not make this clear enough. By placing the visual aid on a separate page (in a landscape fashion), this was improved. It meant that the horizontal line could be significantly lengthened and be large enough to allow the verbal anchors to clearly be associated with the rating-scale endpoints.

Modification: Consequently, the IRSPr3 has the rating-scale visual aid on a separate sheet of paper accompanied by the title ‘Opinion Scale’ (Figure 4. 13). This would mean that the third version of the IRSP would have the respondents follow the instructions and anchor their rating-scale on a

separate sheet of paper. Given all the above considerations, the following instruction was added to the IRSPr3:

### **Defining your own opinion scale**

Here you will need to define your own opinion scale, which you will use to rate the statements you read previously. Please have the 'Opinion Scale' sheet in front of you.

There was an instruction within the IRSPr2 that was an error, which confused some of the respondents. Line 5 of the IRSPr2 instruction sheet read "Please think of a word (adjective) that best describes..." This word "adjective" should have said 'adverb'. This would explain some of the confusion on this point as demonstrated in Interview 5 below:

“\*Respondent 5: What kind of word should I write there? [referring to point (a) on the scale line]

\*Respondent 5: What sort of word do you want me to write there? Like adjective – ‘doing’ word - like what?”  
[Interview 5; 83-91]

“\*Respondent 7: I think the word ‘adjective’ threw me off a bit.”  
[Interview 7; 341-342]

On completion of round two of the qualitative interviews, it was considered whether removing the word ‘adjective’ and replacing it with the correct term ‘adverb’, would still confuse some respondents. Whilst words like ‘strongly’, ‘completely’, and ‘totally’ are adverbs, it was questioned whether, generally, people would know the meaning of the word ‘adverb’. Whether it would be best for respondents to be asked simply to place a ‘word’ in front of ‘agree’, to form the phrase that represents the most they could possible agree with a statement, was considered. Alternatively they could be asked to “place an adverb in front of the word ‘agree’”. Whilst asking them to think of an adverb

to place in front of the word ‘agree’ is *technically* correct, it was thought possible that this would have over-complicated the simplicity of the instruction for those that were somewhat unsure about the definition of the term ‘adverb’.

The decision was taken to spot-test this briefly by asking a number of students whether they knew what ‘adverb’ meant. Approximately 25 students were approached in and around the central building of Leeds University campus, and they were asked whether they understood what the term ‘adverb’ meant. Around ten attempted the answer, and only 2 of these were correct in their understanding of which words are ‘adverbs’. The other 15 simply ‘had no idea’. Consequently, it was deemed prudent to refrain from including the term ‘adverb’ within the instruction, as it seemed likely to provoke some confusion.

Modification: Subsequently the wording of this instruction (D in Figure 4. 9);

Please think of a word (adjective) that best describes the most you could possibly agree with a statement. Please write this word on the above line in the space labelled (a).

was modified into:

Now think of a word to put next to ‘agree’ that would describe the most you could possibly agree with a statement. Please write this word clearly in the space labelled (a).

The phrase ‘the most you could possible agree’ was underlined to highlight its importance to the respondent.

Modification: It was thought useful to inform respondents that the ‘agree verbal anchor’

they defined would in future be referred to as ‘(a)’. This was done in

order to prevent too much repetition of phrases in subsequent instructions. The sentence, shown in bold, was therefore added.

Now think of a word to put next to 'agree' that would describe the most you could possibly agree with a statement. Please write this word clearly in the space labelled (a). **The most you could possibly agree with a statement will now be referred to as (a).**

Modification: When examining the next instruction that followed (E in Figure 4. 9):

If your opinion towards something was statement (a), think about how many steps away from 0 (the neutral opinion) would best represent your view? In other words, please assign a number which you feel would best represent the level of agreement you have described in statement (a). Please write this number in the box labelled (b).

It was quite clear that this could be improved significantly, and simplified. It was decided that the first sentence:

If your opinion towards something was statement (a), think about how many steps away from 0 (the neutral opinion) would best represent your view?

could be replaced with the following:

Now think about how many steps you have between feeling 'neutral' and feeling (a) towards a statement.

Modification: It was also considered whether providing an example, of what was meant,

would help respondents understand the task (ideally without 'influencing' them). Therefore the next phrase was added:

For example if a person had one stage between 'neutral' and (a), then feeling (a) would be their 2<sup>nd</sup> step. This person would write a 2 in the box labelled (b).

Modification: It was thought crucial that the next sentence reinforce the attention back to the personal experience of the respondent. So the subsequent sentence was added to the instruction:

How many stages of agreement do you feel you have?

Modification: In order to help encourage respondents to make annotations to the visual aid to help them visualise their scale, the following instruction was added,

To make it easier, you can mark the steps on the line to help you think. Write the number that corresponds to (a) in box (b).

Modification: The changes applied to the instruction calling for the ‘negative verbal anchor’ to be defined, mirrored the changes applied above to the ‘positive verbal anchor’. Therefore the following instruction (F in Figure 4. 9);

Please think of a word (adjective) that best describes the most you could possibly disagree with a statement. This word can be the same as or different to the word you wrote in (a). Please write this word on the above line in the space labelled (c).

was modified into:

Now think of a word to put next to ‘disagree’ that would describe the most you could possibly disagree with a statement. This word can be the same as, or different to the one you thought of before. Please write this word clearly in the space labelled (c). The most you could possibly disagree with a statement will now be referred to as (c).

Modification: Changes made to the instruction asking respondents to define the ‘negative numerical anchor’ mirrored those changes described above to



the instructions for the ‘positive numerical anchor’. The only difference was that there was no sentence giving an ‘example’, as there should be no need for a second example. Additionally, the sentence reminding the respondent that the number had to be negative, was removed. This was done in order to see whether the instruction was really necessary (i.e. whether stating that the “number had to be negative” was pointing out the obvious). The new instruction was phrased in the following manner:

Now think about how many steps you have between feeling ‘neutral’ and feeling (c) towards a statement. How many stages of disagreement do you feel you have? If you like you can mark the steps on the line to help you think. Write the number that corresponds to (c) in box (d).

The final instruction in the original IRSPr2 read as follows (H in Figure 4. 9):

You now have your own personally-defined measurement scale. Please use this scale to show how strongly you agree or disagree with the subsequent statements. Please do so by writing the number from your scale which best corresponds with your opinion in the spaces given next to each statement.

Modification: The above instruction was changed in the following ways:

- The phrase ‘personally-defined measurement scale’ was replaced with ‘personally-defined opinion scale’ to reflect the insights gained about the word ‘opinion’ being more accessible to respondents (as explained previously).
- The phrase ‘how strongly’ was replaced with ‘how much’ to remove bias in the instructions, given that ‘strongly’ is an adverb a respondent could choose to use as a verbal anchor.
- The word ‘subsequent’ was removed, just to keep the instruction simple.

- ‘which best corresponds with your opinion’ was changed to ‘which best represents your opinion’.
- The word ‘spaces’ was changed to ‘boxes’, as it is more specific.

As such, the newly modified instruction, read as follows:

You now have your own personally-defined opinion scale. Please use this scale to show how much you agree or disagree with the statements. Please do so by writing the number from your scale which best represents your opinion in the boxes next to each statement.

#### 4.2.6.3 *Potential improvements*

##### *List of verbal anchors*

Interviewee 4 chose ‘10’ for the ‘agree pole’ but chose ‘-8’ for the ‘disagree pole’, resulting in an IRS of -8←0→10. He was the first to define an imbalanced IRS (in the sense that there are more intervals on his ‘agree pole’ than on his ‘disagree pole’). The interesting thing with this respondent is that he said he sees ‘disagreeing’ and ‘agreeing’ as being mirror opposites of one another (bipolar), however, he rated them -8 and 10. When asked why despite seeing them as mirror opposites he assigned different numbers, he explained that it was because he felt he could *agree* with something on *more* levels than he could *disagree* with something. However, it was considered likely that his choice of adverbs ('emphatically' and 'absolutely'), affected his choice of different numbers. He stated that 'absolutely' was the most he could disagree with something, for him, but that other people may have other words to indicate a stronger level of disagreement,

"I felt that for me ‘emphatically’ is the highest you can go, but as I was saying before for me ‘absolutely’ – there’s probably other words other people would use which could be *more*, so that’s why – well for me that the most..."

[Interview 4: 141-144].

His admission that others might not view it as a ‘maximum’, was considered a clue, in that he may have been unable to think of a stronger adverb (for disagreeing). Perhaps his choice of the word ‘absolutely’ led to a restriction in what he felt he could assign to it *numerically*. In this way he may have ‘downgraded’ it from the strength he gave to the word ‘emphatically’. Perhaps he was unwilling to state that he could not think of a stronger adverb and took refuge in saying his view of agreeing and disagreeing is imbalanced – which contradicts his earlier statement, that he sees them as ‘mirror opposites’. He is someone who appeared to be confident and a somewhat proud individual. Interestingly, he also said that he encountered no difficulties with the exercise at all, and did not criticise any part of the instructions when invited to do so. So it was thought that perhaps he is someone who likes to appear very intelligent and someone of strong convictions, who would not naturally have chosen to pick a numerically imbalanced scale if it were not for the mismatch in the strength of the adverbs chosen.

The link between the strength-of-adverbs-chosen and the numerical anchors assigned, reappeared in Interview 7. Interviewee 7 indicated that in choosing the word ‘strongly’ for both anchors, he was defining a typical rating-scale. However, he stated that strongly (dis)agree was not the most he could (dis)agree with something, and that when he assigned 4 and -4 to the endpoints, he was aware of this fact;

\*Interviewer; Right, what about the number 4. Why the number 4?

\*Respondent 7; Umm...again I put ‘strongly’, ‘absolutely’ would’ve been a 5 –

[Interview 7: 212-215]

Here, he indicates quite clearly that had he chosen the word ‘absolutely’ he would have assigned the number ‘5’ to it. Therefore, it was clearly quite important that respondents

choose verbal anchors that are personally relevant and that spanned their entire agreement-disagreement spectrum, as it seemed to strongly affect their numerical anchoring.

It was considered whether respondents might benefit from having a list of adverbs provided. This would enable respondents to inspect a list of adverbs (thus saving them the task of thinking of them), but still allow them to personalise their IRS through the process of choosing which adverbs to use and subsequently assigning their numerical anchors.

At this stage no modification was made for several reasons: (a) It had the potential to lengthen the IRSP process substantially if respondents had to inspect a list of adverbs before making a choice. This might be quite inefficient; (b) There would be bias introduced as the researcher would be choosing the adverbs to include in the list; (c) It might have the potential to frustrate respondents. For these reasons, it was decided that such an avenue would only be pursued if more evidence presented itself. At this stage, it was clear that the instruction wording concerning numerical and verbal anchors could still be improved. A note was made to probe for feedback on ‘List of Adverbs’ and whether or not interviewees would have found them helpful.

#### *An electronic IRSP*

Respondent 3 was discussing the flow of the IRSP instructions, and when asked about some of the ways in which the instructions could have been improved, she stated “colour coding would have made it easier for me.” [Interview 3: 100]. When further probed she gave the following example:

“\*Respondent 3: If the ‘a’ here was like red, and the ‘a’ here was red [pointing at the (a) on the scale line and (a) in the instructions]”

[Interview 3; 103-105]

She indicated that she is more of a ‘visual’ person and that matching each specific instruction in the text against the rating-scale visual aid using a colour coded system would have helped her. For example, colouring red the sentence instructing the respondent to verbally anchor their maximum agreement, and colouring red the gap next to ‘agree’ (in the visual aid). She was asked whether she felt this modification would have improved simply the *speed* with which it took her to complete the whole process or whether it would have augmented her *understanding* of what was being asked of her. She indicated that the colour-coding would have merely increased the *speed* with which she completed the process. Whilst this was an interesting suggestion, it was felt that if taken literally, this would produce a more confusing version of the IRSP for respondents. It would have been risky to assume that all respondents would take well to this colour-coding system. It would potentially make the instructions look even more ‘busy’, especially for respondents who are not as ‘visual’ in their logical reasoning as Interviewee 3. Given that implementing this change would not improve the respondents’ *understanding* of the instructions, it was decided that it was not necessary.

Nevertheless, there were still insights to be gained from considering Interviewee 3’s *reason* for her colour-coding suggestion. Clearly her idea was born out of a problem; the problem being that she did not feel as though the rating-scale visual aid paired well with the instructions. In other words, there might be a better way of pairing up each instruction with its counterpart in the rating-scale visual aid. At this stage the instructions contained a lot of (a)s, (b)s, (c)s and (d)s and so did the visual aid. This was

necessary in order to show the respondent which part of the visual aid each instruction was referring to. Whilst the IRSP remained a paper-based process (i.e. a paper questionnaire) this would be difficult to avoid, or improve upon. This raised the question of whether this Individualised Rating-Scale Procedure could be developed much further on paper. With the IRSP becoming a computer-based process, it would introduce a dynamism and simplicity that could not be achieved on paper. For example, the instructions could be shown one at a time with their counterpart visual aid, and the respondent could be shown a final visual aid of the finished product (i.e. their IRS with all their intervals 'drawn on' for them). This would greatly simplify the instruction process and generate a much better picture of the individualised rating-scales than could be achieved on paper. However, before developing a computerised IRSP, a final paper-based check of the instruction-wording was necessary.

This summarises the changes made to the instructions of the paper-based IRSP, resulting in the IRSPr3 instruction sheets (Figure 4. 12 and Figure 4. 13). These instruction sheets were used to proceed onto round three of the interviews.

	<u>Instructions</u>
A →	<p>Now think about the statements you have just read.</p> <ul style="list-style-type: none"> <li>• Think about whether or not you agree or disagree with each statement.</li> <li>• Also think about <u>how much</u> you agree or disagree with each statement.</li> <li>• If you have no opinion on a statement then your position is neutral. This means that you don't agree with it, and you don't disagree with it.</li> </ul>
B →	<p><b>Defining your own opinion scale</b></p> <p>Here you will need to define your own opinion scale, which you will use to rate the statements you read previously. Please have the 'Opinion Scale' sheet in front of you.</p>
C →	<p>The number 0 on the scale, will represent your neutral position (i.e. no opinion). This has been labelled on the line. This means that if you didn't agree or disagree with a statement it would be rated 0.</p>
D →	<p>Now think of a word to put next to 'agree' that would describe <u>the most you could possibly agree</u> with a statement. Please write this word clearly in the space labelled (a). The most you could possibly agree with a statement will now be referred to as (a).</p>
E →	<p>Now think about how many steps you have between feeling 'neutral' and feeling (a) towards a statement.</p> <p style="padding-left: 40px;">For example if a person had one stage between 'neutral' and (a), then feeling (a) would be their 2<sup>nd</sup> step. This person would write a 2 in the box labelled (b).</p> <p>How many stages of agreement do you feel you have? To make it easier, you can mark the steps on the line to help you think. Write the number that corresponds to (a) in box (b).</p>
F →	<p>Now think of a word to put next to 'disagree' that would describe <u>the most you could possibly disagree</u> with a statement. This word can be the same as, or different to the one you thought of before. Please write this word clearly in the space labelled (c). The most you could possibly disagree with a statement will now be referred to as (c).</p>
G →	<p>Now think about how many steps you have between feeling 'neutral' and feeling (c) towards a statement. How many stages of disagreement do you feel you have? If you like you can mark the steps on the line to help you think. Write the number that corresponds to (c) in box (d).</p>
H →	<p>You now have your own personally-defined opinion scale. Please use this scale to show how much you agree or disagree with the statements. Please do so by writing the number from your scale which best represents your opinion in the boxes next to each statement.</p>

**Figure 4. 12 IRSP Instruction Sheet IRSPr3 Part 1: Interviewee 8**

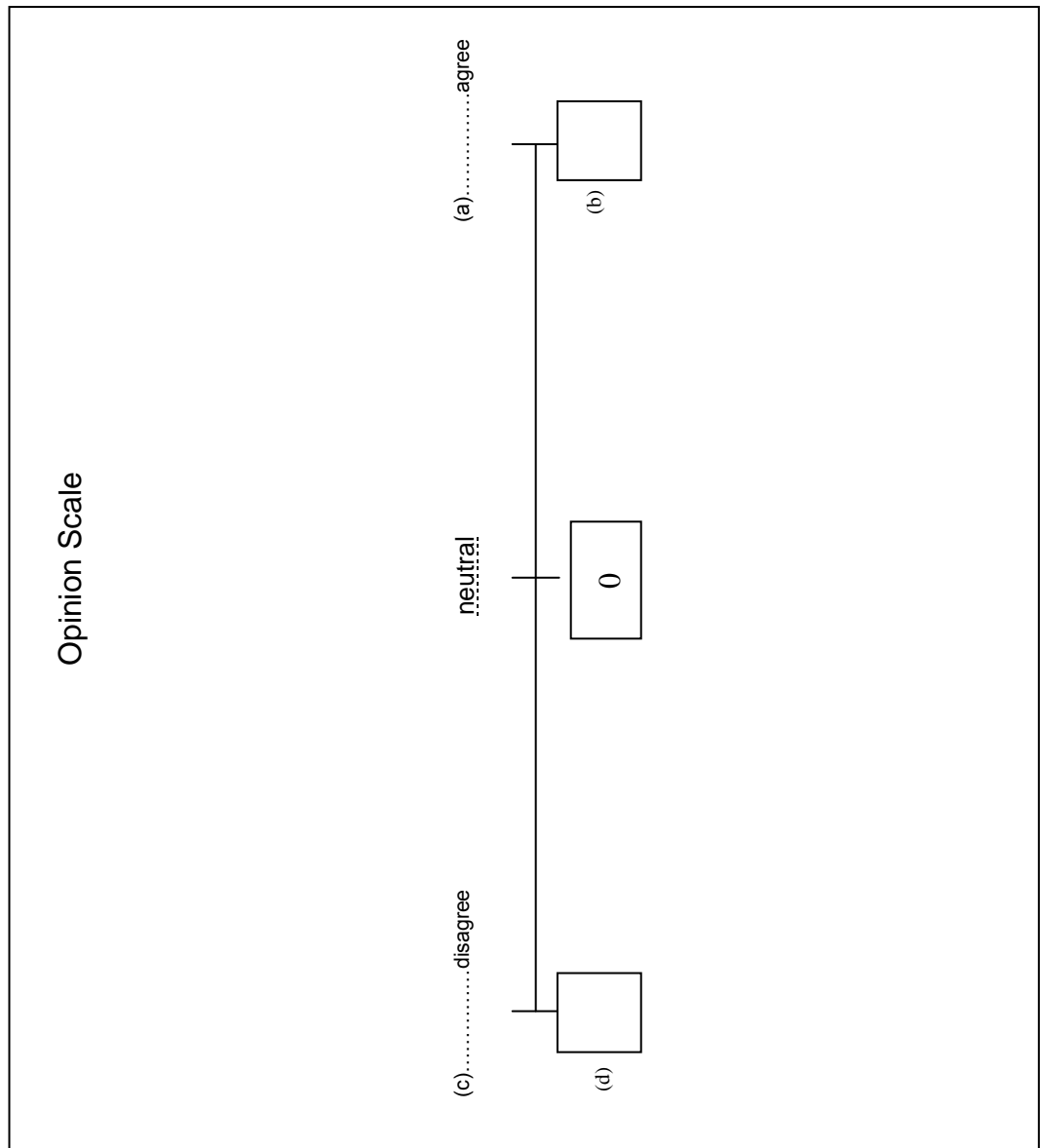


Figure 4. 13 IRSP Instruction Sheet IRSPr3 Part 2: Interviewee 8

**4.2.7 Round 3 – Interview 8**

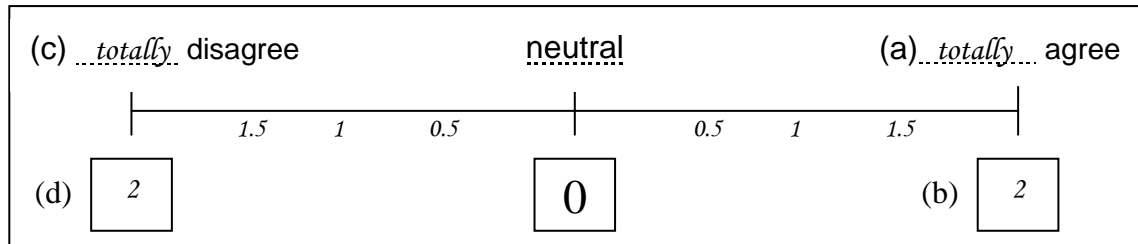
Interview 8 provided enough insight to warrant modifying the IRSP before proceeding any further, which is why this round only consisted of one interview. In addition, the interview came to a premature end because the interviewee was unable to continue due to an unforeseen interruption. She was called away after completing the entire IRSP. As such, most of the interviewer’s questions had been posed.



#### 4.2.7.1 Key finding

##### *Individualised Rating-Scales (IRSs) chosen*

Interviewee 8 defined the following rating-scale, including the annotations drawn on the line. The original IRS drawn by Interviewee 8 had numerical anchors from -2 to 2, and she sketched 1.5 and 0.5 on the horizontal line.



**Figure 4. 14 Interviewee 8: IRS Chosen**

Note that Interviewee 8 did not put a minus sign in front of the ‘2’ written in box (d), as is shown.

Interviewee 8 took approximately 26 seconds to read Greenleaf’s statements, 2 minutes to read and execute the IRSP and 52 seconds to rate all the statements using her IRS. Total time taken (not including interruptions) was 3 minutes and 30 seconds. This would indicate that the IRSP is still being completed within a reasonable time-frame.

#### 4.2.7.2 Key modifications

##### *The biasing effect of the ‘example’*

It became clear that the example (for numerical anchoring) included in IRSPr3 appeared to be biasing the Interviewee. Recall that this phrase was included in IRSPr3 (E in Figure 4. 12):

For example if a person had one stage between ‘neutral’ and (a), then feeling (a) would be their 2<sup>nd</sup> step. This person would write a 2 in the box labelled (b).

It was included to help respondents understand the task, ideally without influencing their selection of numerical anchors. Given Interviewee 8 had chosen the number ‘2’ for both endpoints, it became all the more important to determine her decision process for doing so to ensure the *example* had not biased her. The first clue was that she had paused to clarify the following,

“\*Respondent 8: So for these ones, do you just have to decide how many stages there is up until?...that one like 10 or whatever [she points to the positive blank on the scale]?”

[...]

\*Respondent 8: I’m going to put a 2 in there.”

[Interview8: 42-45, 61-62]

Interestingly, her mention of the number ‘10’ shows a continued attraction to this number with numerical anchoring. It appeared as though her decision to place a ‘2’ in the box occurred immediately after she read the *example*. Whilst it is advantageous that the *example* appeared to influence her into choosing a shorter rating-scale, it was worrying that she chose the exact same number that was used in the *example*.

Just before numerically anchoring her absolute disagreement, the following discussion took place,

“\*Respondent 8: I don’t have to put a 2 in this box? [respondent points to disagreement numerical endpoint]

\*Interviewer: Why do you think that?

\*Respondent 8: [respondent laughs] Just assuming...seeing as I had a 2 in that one [respondent points to agreement numerical endpoint]

\*Interviewer: You can put whatever you like in that box.

\*Respondent 8: [respondent giggles] ...right...

\*Interviewer: Why did you think you don’t have to put a 2 in the box?

\*Respondent 8: I don’t know [respondent laughs]

\*Interviewer: [interviewer laughs]

\*Respondent 8: I'll just put a 2, and then I can just decide on a scale all the way up to here. Yeh."

[Interview 8: 96-116]

It seemed clear that the absence of an *example* for the numerical anchoring of her absolute disagreement left her unsure as to what number she should write in box (d). This continued to suggest that the earlier *example* had had a biasing effect on her actions.

As shown in Figure 4. 14, she had annotated the line either side of neutral with '0.5', '1' and '1.5'. Thus, she had three stages between neutral and both her endpoints.

“\*Interviewer: Right, that’s interesting. I’m going to stop you there. Now you’ve put 0 in the middle, 2 on one side and 2 on the other side.

\*Respondent 8: Yep.

\*Interviewer: And you’ve split it up so that it goes up in 0.5 levels...so you’ve got 0.5, 1 and 1.5. What do those mean to you? You’ve got three levels in between there, so what do they mean to you?

\*Respondent 8: That would be like ‘maybe’ [pointing at 0.5] then that would be like ‘more-so’ [pointing at 1], and then ‘even more’ [pointing at 1.5] and then ‘totally agree’ [pointing at 2]”

[Interview 8: 117-129]

Given the discussion, it became all the more important to probe further into why she had chosen the number ‘2’, and not ‘4’ (given she indicated she had three ‘in-between’ stages).

“\*Interviewer: Totally, right ok. Why did you not do let’s say – because I’m counting, here you’ve got 1, 2, 3 and then ‘totally agree’ would be your fourth level – so why did you not put 4 in the box? Why did you put 2?

\*Respondent 8: No idea [laughs]”

[Interview 8: 130-135]

The discussion continued and simply confirmed that the inclusion of the *example* did in fact have a biasing effect. Interviewee 8 did not acknowledge this directly. However there was enough observational evidence to make this deduction.

Modification: The *example* was removed from the IRSP instructions, resulting in a new instruction sheet for use in Round 4 of the interviews, IRSPr4 (see Figure 4. 15).

Additionally, Interviewee 8 suggested that it might be useful for the instructions to say “go up in whole numbers” [Interview 8: 169], as a means of making sure others do not use decimals. This was considered to be a very useful suggestion.

Modification: As such, a simple modification was made, with the following sentence being added:

Think about it step by step, in terms of whole numbers.

Below you can see the original instruction (E in Figure 4. 12) and the new sentence. The new sentence has been underlined, to illustrate its positioning.

Now think about how many steps you have between feeling ‘neutral’ and feeling (a) towards a statement. How many stages of agreement do you feel you have? Think about it step by step, in terms of whole numbers. To make it easier, you can mark the steps on the line to help you think. Write the number that corresponds with feeling (a), in box (b).

The amendment was also mirrored for the ‘disagreement’ side of the continuum, namely:

Now think about how many steps you have between feeling ‘neutral’ and feeling (c) towards a statement. How many stages of disagreement do you feel you have? Think about it step by step, in terms of whole numbers. If you like you can mark the steps on the line to help you think. Write the number that corresponds with feeling (c), in box (d).

*Confusion over the 'sign' of disagreement*

The fact that there was no part of the instruction that stated the number in box (d) had to be negative, seemed to cause a bit of confusion. Interviewee 8 asked,

“\*Respondent 8: It’s just that how will you know if I agree or disagree?

[...]

\*Respondent 8: Do you want me to just put a ‘d’ or ‘b’ -”

[Interview 8: 191-195]

Here she was asking whether, when rating the statements, she was expected to write ‘2d’ for ‘disagree’, and ‘2b’ for ‘agree’ (i.e. the number ‘2’ and box (b)). She later suggested a minus or a plus sign.

Modification: It seemed wise to bring back a sentence stating the ‘disagreement’ number had to be negative, to avoid any confusion. So the following sentence was added to the end of the paragraph for the ‘disagreement’ side of the continuum:

This must be a negative number.

This summarises the changes made to the instruction sheets. No changes were made to the visual aid, as no problems were evident. The IRSPr4 instruction sheets were used to proceed onto Round 4 of the interviews, shown in Figure 4. 15.

	<u>Instructions</u>
A →	Now think about the statements you have just read. <ul style="list-style-type: none"> <li>• Think about whether or not you agree or disagree with each statement.</li> <li>• Also think about <u>how much</u> you agree or disagree with each statement.</li> <li>• If you have no opinion on a statement then your position is neutral. This means that you don't agree with it, and you don't disagree with it.</li> </ul>
B →	<b>Defining your own opinion scale</b> Here you will need to define your own opinion scale, which you will use to rate the statements you read previously. Please have the 'Opinion Scale' sheet in front of you.
C →	The number 0 on the scale, will represent your neutral position (i.e. no opinion). This has been labelled on the line. This means that if you didn't agree or disagree with a statement it would be rated 0.
D →	Now think of a word to put next to 'agree' that would describe <u>the most you could possibly agree</u> with a statement. Please write this word clearly in the space labelled (a). The most you could possibly agree with a statement will now be referred to as (a).
E →	Now think about how many steps you have between feeling 'neutral' and feeling (a) towards a statement. How many stages of agreement do you feel you have? Think about it step by step, in terms of whole numbers. To make it easier, you can mark the steps on the line to help you think. Write the number that corresponds with feeling (a), in box (b).
F →	Now think of a word to put next to 'disagree' that would describe <u>the most you could possibly disagree</u> with a statement. This word can be the same as, or different to the one you thought of before. Please write this word clearly in the space labelled (c). The most you could possibly disagree with a statement will now be referred to as (c).
G →	Now think about how many steps you have between feeling 'neutral' and feeling (c) towards a statement. How many stages of disagreement do you feel you have? Think about it step by step, in terms of whole numbers. If you like you can mark the steps on the line to help you think. Write the number that corresponds with feeling (c), in box (d). This must be a <u>negative</u> number.
H →	You now have your own personally-defined opinion scale. Please use this scale to show how much you agree or disagree with the statements. Please do so by writing the number from your scale which best represents your opinion in the boxes next to each statement.

**Figure 4. 15 IRSP Instruction Sheet IRSPr4 Part 1: Interviewees 9-13**

**4.2.8 Round 4 – Interviews 9-13**

*4.2.8.1 Key findings*

*Individualised Rating-Scales (IRSs) chosen*

The IRSs chosen by interviewees 9-13 are summarised here in Table 4. 4.

**Table 4. 4 Interviewees’ 9-13: Numerical and Verbal Endpoints Chosen**

Interview	Verbal		Numeric		Changes Verbal		Changes Numeric	
	+	-	+	-	+	-	+	-
9	Definitely	Really	10	-10	None	None	3	-3
10	Strongly Experience that always	Strongly	3	-3	None	None	None	None
11	Definitely	Detest to do	3	-3	Totally	Totally	None	None
12	Definitely	Completely Very	5	-5	None	None	None	None
13	Absolutely	strongly	4	-4	None	None	None	None

At this stage, less people used the number ‘10’. Only Interviewee 9 elected to use it, but later expressed a desire to change it to -3←0→3. Interviewee 11 interpreted the verbal anchoring process as an area where he had to indicate the ‘frequency’ or ‘infrequency’ with which he *felt* each of Greenleaf’s statements. After realising that he was simply expressing the strength to which he agreed/disagreed with the attitudes expressed, he indicated that he was treating both endpoints as ‘totally agree’ and ‘totally disagree’. He was observed ‘skim-reading’ the instructions, which may have led to his misinterpretation.

*IRSP execution times*

Table 4. 5 shows the time it took each interviewee to carry out the instructions. As can be seen, the entire exercise took on average around 5 minutes and 45 seconds. This demonstrates that the average time to complete the process continued to be reasonable.

**Table 4. 5 Interviewees' 9-13 Exercise Completion Times**

Interview no.	Greenleaf statements		IRSP (Reading and Executing)	Total (mins)
	Reading time (secs)	Rating time (secs)	Time (secs)	
9	46	138	134	5.30
10	10	86	80	2.93
11	49	159	269	7.95
12	40	98	142	4.67
13	49	189	146	6.40
<b>Avg time</b>	<b>38.8</b>	<b>134</b>	<b>154.2</b>	<b>5.45</b>

*The need for IRSP software*

At this stage it was quite clear that, assuming feasibility, the dynamic and personalised premise of the IRSP could benefit tremendously from electronic development. Therefore, the modifications made as a result of findings from this round (interviews 9-13) were done with the electronic IRSP in mind.

*List of adverbs*

On analysing the data from Round 2 (interviews 3-7) it was noted that a list of adverbs might potentially help respondents with their selection of verbal endpoints. As such, this idea was put forward to the interviewees in Round 4 (note: this would also have been done in Round 3 had interview 8 not ended prematurely).



**Table 4. 6 Overview of Respondents' Desire for a List of Adverbs**

Interview: Lines	Question Posed	Response
[9: 92-103]	“*Interviewer Ok. Umm...if I'd have thought of some other words – let's say you'd have had a list of other words in front of you – [...] – umm to help you...adverbs, such as ‘totally’, ‘absolutely’, ‘emphatically’... would any of those words have aided you? Or would you have used those words instead of any of the words you've put? Or are you happier with the words you've put?”	“*Respondent 9 I'm happy with mine.”
[10: 75-80]	“*Interviewer Right, that's really good. And um, just another question on that, do you feel that you needed a list of adverbs to help you? Or were you quite happy thinking of those words yourself?”	“*Respondent 10 I was quite happy.”
[11]	Although Interviewee 11 was not asked this question during the taped interview, it was mentioned to him afterwards. He indicated that he would have preferred to come up with the adverbs on his own, and that if the instructions were made a little clearer, he would have thought of them first time round.	
[12: 390-397]	“*Interviewer Yeah. Ok, so you didn't feel as though you needed a list of adverbs to help you? Like things like ‘absolutely’, ‘totally’ etc”	“*Respondent 12 No. No. I think because it's such a personal thing, it's your own scale so it needs to be... <i>you</i> need to define it yourself, so using whatever vocabulary you have...and your own understanding of what those words mean to <i>you</i> .”
[13: 116-121]	“*Interviewer Did you feel that you needed the aid of a list of adverbs to help you, for example, let's say with ‘disagree’ would you have wanted to put another word in front of it but you couldn't think of one?”	“*Respondent 13 No.”

All interviewees were happy not to have a list supplied. Interestingly, some of the interviewees appeared to have a strange sense of achievement with the rating-scale they defined. Not only did all of them feel that they would prefer to choose their own adverbs without the aid of a list, but Interviewees 12 and 13 shed some interesting light on the issue. Interviewee 12 kept on emphasising that ‘*your* words’ should be used to create ‘*your own* scale’. It was quite clear that she felt that she would have been ‘contaminated’ had she seen a list of adverbs. She felt that whilst some might have a more expansive knowledge of adverbs than others, *everyone* is capable of choosing

verbal anchors that are personally meaningful to them when describing their maximum agreement/disagreement. She believed that a list of adverbs would have detracted from the personalised nature of the rating-scale. Interviewee 13 is an exciting case due to the fact that she appeared to personify Interviewee 12's comments about personally meaningful endpoints for a personalised rating-scale. Interviewee 13 chose verbal endpoints 'very strongly disagree' and 'absolutely agree'. When asked why she chose to use different words (the routine interviewer prompt for differing verbal anchors), she responded with,

“\*Respondent 13: Um, to be honest I thought to myself I couldn't say 'absolutely disagree' because it sounds stupid.”

[Interview 13: 74-76]

In other words, she first defined her maximum agreement as 'absolutely agree', and had considered using the same adverb for 'disagree' and did not because "it sounds stupid". Obviously, this was a rather curious and humorous response and so she was probed further.

\*Interviewer: Why does it sound stupid?

\*Respondent 13: Um, 'absolutely disagree'...don't know it just wasn't right.

\*Interviewer: So you thought about –

\*Respondent 13: 'Absolutely agree' with something, 'very strongly disagree' with something. I think in general speech that's the kind of thing I'd use. I would – oh don't worry.

\*Interviewer: No, no, keep talking, that's very interesting.

\*Respondent 13: I would say 'very strongly disagree', but for some reason I'd say 'absolutely agree' with that instead of 'very strongly'.

\*Interviewer: So you felt that in speech, like, from experience...you just don't feel you'd say it naturally, 'absolutely disagree', you wouldn't say that naturally...

\*Respondent 13: Yeah. See 'absolutely', is 'I'm agreeing with you'. 'Absolutely'.

\*Interviewer: Yeah that's very true, because it is. If someone says 'do you like this ice-cream?' it's 'absolutely'. And people would assume that you're saying you agree.

\*Respondent 13: But in terms of...if you disagree with something – I'd very rarely say 'absolutely not'."

[Interview 13: 77-104]

Interviewee 13 indicated that the word 'absolutely', when taken on its own, is a word commonly used to express 'agreement', and that she would not naturally use it in conversation to express 'disagreement'. This demonstrates how her chosen endpoints appear to be personally meaningful to her. Even more interestingly, she revealed that she was picturing what she would say 'naturally', in other words she may have imagined herself agreeing/disagreeing with someone and recalling how she would *naturally* express her agreement/disagreement in conversation. This is a clue as to how the IRSP could be further improved; through encouraging respondents to picture themselves in an agreeing/disagreeing scenario, in order to help them access personally meaningful verbal anchors.

Thus, the investigation over the inclusion of a list of adverbs was dropped. It was decided that there could be a better way of refining the instructions for the verbal anchors so that respondents could be encouraged to choose verbal labels that they would naturally use in speech (i.e. in expressing their absolute agreement/disagreement with something in natural conversation). It was considered that this might enable respondents to connect better with the words they choose, and in turn assign more meaningful numerical values.

Modification: The following instruction (D in Figure 4. 15);

Now think of a word to put next to 'agree' that would describe the most you could possibly agree with a statement. Please write this word clearly in the

space labelled (a). The most you could possibly agree with a statement will now be referred to as (a).

was changed into the following:

Now think of a word to put next to 'agree' that would describe the most you could possibly agree with a statement.

If it helps, try and picture yourself talking to someone, and they said something that you agree with as much as you possibly could. Picture yourself responding to this person by saying "I ----- agree".

Please write your word clearly in the blue box.

There was no need to refer to the box as '(a)' because the computer program would be able to simplify the visual aid presented to respondents. In addition, respondents' chosen endpoints would no longer need to be called '(a)' and '(c)' when referenced, because the program would be able to weave the respondents' inputs into subsequent instructions, through the code. In other words, the instructions would be dynamic.

Additionally, the same changes were replicated for the verbal-anchoring instructions for disagreement. So the following instruction (F in Figure 4. 15);

Now think of a word to put next to 'disagree' that would describe the most you could possibly disagree with a statement. This word can be the same as, or different to the one you thought of before. Please write this word clearly in the space labelled (c). The most you could possibly disagree with a statement will now be referred to as (c).

was changed into:

Now think of a word to put next to 'disagree' that would describe the most you could possibly disagree with a statement.

If it helps, try and picture yourself talking to someone, and they said something that you disagree with as much as you possibly could. Picture yourself responding to this person by saying "I ----- disagree".

Please write your word clearly in the blue box.

It was decided, due to the conceptually meaningful cue, that it was no longer necessary to state that the word chosen could be the same as or different to that chosen previously. So, that statement was not kept in the new version of the instruction.

*The mystery attraction of '±10' revisited*

Upon completion of Interview 13, the interviewee made some enquiries about the research and was particularly curious about the numerical endpoints chosen by others. It was mentioned to her that in previous interview rounds respondents had a tendency to be attracted to  $\pm 10$ . She immediately commented on this, stating that she could understand why their might be this tendency, explaining her reasons. She was asked if she would mind if the tape recorder were switched back on so that some of her insight could be captured. After giving her permission, she restated some of her thoughts.

Below is an extract:

\*Respondent 13: I was just suggesting that perhaps, people think to use a scale of 1 to 10 and -1 to -10, is because as a child that's the first 10 numbers that you're taught, and so you're used to scale everything in terms of 1 to 10.

\*Interviewer: Yeah...Did that pop into your head at any point, the *want* to put the number 10?

\*Respondent 13: Um...

\*Interviewer: Or it didn't?

\*Respondent 13: No it didn't actually, having said that, just – perhaps it did, but I thought to myself 'what is a reasonable scale for this?' i.e. when you were asking, like, why did I choose 1 to 4? Because –

\*Interviewer: You also had this conversation around a month ago.

\*Respondent 13: [laughs] Yeah. There's different...

\*Interviewer: Different ways of levelling it?

\*Respondent 13: Yeah, it's a good – you can differentiate between 1 and 2, on a 1 to 4 as opposed to on a 1 to 10 with 7 and 8.”

[Interview 13: 283-306]

Earlier in the interview, it was established that Interviewee 13, coincidentally, had had a conversation with friends about four weeks previously where she had discussed how many levels of agreement/disagreement she has. She had come to the conclusion that she has four levels (of both agreeing and disagreeing). As such, when asked to numerically anchor her rating-scale she automatically chose -4 and 4. In the above extract she insinuates that she may have considered using  $\pm 10$  if she had not already had this discussion, but that she might still have theorised that it is far easier to distinguish between intervals on a rating-scale of  $\pm 4$  than on one of  $\pm 10$ . In other words she was saying that, had she not had that discussion with her friends, although she might have considered the use of '10', she probably would have come to the conclusion that her rating-scale had too many intervals and therefore not chosen it. Interviewee 13 went on to say,

“\*Respondent 13: But perhaps some people use 10 because –

\*Interviewer: Like you were saying, with learning it –

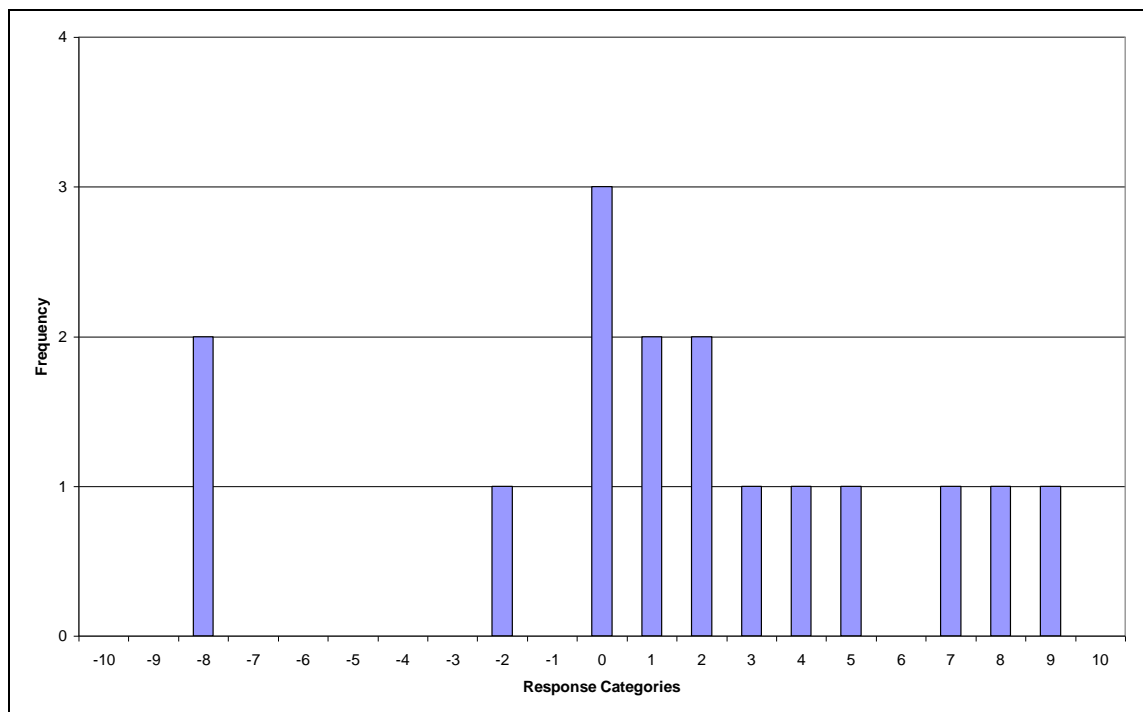
\*Respondent 13: At a very young age. You learn to count to 10 then you learn to count to 20, so everything's done from 10.”

[Interview 13: 309-315]

In short, her point was that the number '10' seems to be this natural marker, and that as children we learn to count first up to ten, and then we learn to count in sets of ten. Ten was also the easiest of the times tables, and the easiest number to think about when dealing with more abstract mathematics like fractions or percentages. In addition, she was saying that it seems to be the natural marker when evaluating things, like “what would you give it/him/her out of 10”. Her comments appear to complement the

explanations given by Interviewees 2, 3, 5 and 6 as to their numerical choices, mentioned earlier in the chapter.

In this fourth round of interviews, Interviewee 9 was the only person, out of five, to choose the ‘mystery  $\pm 10$ ’ for both numerical anchors. This is a big improvement over Round 3 where four out of five interviewees chose  $\pm 10$  for either one or both sides of their IRS. It is clear from Interviewee 9’s spread of responses to Greenleaf’s sixteen items that his IRS probably had too many intervals, as shown in Figure 4. 16.



**Figure 4. 16 Interviewee 9 – Spread of responses**

When examining his verbal responses in relation to the use of his rating-scale, the following was revealed:

“\*Interviewer: [...] I’m going to quickly ask you...if I say ‘university education is very important for success in today’s world’ –

\*Respondent 9: Yep.

\*Interviewer: And I say ‘I like to visit places that are totally different from my home’,-

\*Respondent 9: Yep –

\*Interviewer: Which one of those do you agree with more?

\*Respondent 9: Umm...probably the same.

\*Interviewer: Probably the same? Ok, because here you actually rated them 8 and 9...so you have rated them differently. Which means, yeh it's probably down to the length of the scale –

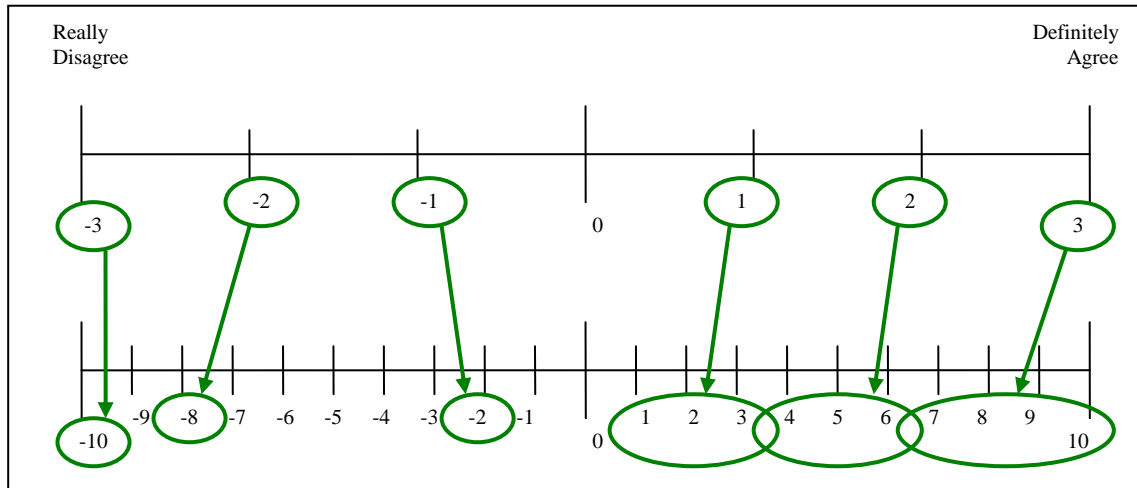
\*Respondent 9: Yes.”

[Interview 9: 211-230]

This suggested that having ten intervals on the agreement side of the pole proved to be too many for this individual. It would seem that not every interval was as distinct as it should be. Interviewee 9 recognised this before the above discussion took place. He indicated that he would change his IRS to  $\pm 3$ . Interestingly, when examining his spread of responses, he only used two intervals on the disagreement side of his IRS and eight intervals from the agreement side. This might give the impression that he suffers from the tendency to acquiesce. However, there could be an alternative explanation for the above spread of responses.

Hui and Triandis (1989) demonstrated how a respondent using a rating-scale of an unsuitable length results in them having to map their ideal rating-scale onto the rating-scale provided by the researcher. They showed how this process of mapping one onto the other is frequently done unevenly, and illustrated how clusters might form. Figure 4.17 is a hypothetical example of how Interviewee 9 may have mapped out his own ideal rating-scale. Whilst his spread of responses might suggest that he has a tendency to acquiesce, an alternative explanation could be that he may have had a more even spread of responses if he had defined his *ideal* IRS. It was therefore important that the IRSP be able to guide respondents into identifying and accessing what, for them, is the most meaningful IRS.





**Figure 4.17 Hypothetical example: Mapping one's ideal IRS onto one with too many intervals**

Whilst the findings continue to support the notion that there is a tendency to choose  $\pm 10$ , the objective is that the IRSP would steer respondents to think more carefully about their numerical choices, to ensure more meaningful intervals. In thinking about how the IRSP could be further improved, to break through this peripheral attraction to  $\pm 10$ , a new idea was conceived. This was referred to as the IRSPv2, and is explained in the later section entitled 'IRSPv2'.

#### *Personally meaningful IRSs*

It was clearly advantageous for the quality of data capture, that respondents were able to define IRSs that were personally meaningful, both in terms of verbal and numerical conceptualisation. Round 4 of interviews showed that the IRSP was proceeding in the right direction, given that most of the interviewees appeared to be defining and using personally meaningful IRSs. Discussions with the interviewees also proved insightful when probing this issue. When asked how they came to choose their numerical anchors, most of the interviewees' explanations evidenced a thought-process that was purposeful

and meaningful. Interviewees 10 and 13 had quite clearly thought about what each of their intervals ‘meant’ to them *before* assigning their numerical endpoints.

[Interview 10: 81-86]

\*Interviewer: You were quite happy. Great. What about the number 3 either side? Why 3?

\*Respondent 10: Just varying degrees of how strongly I would agree or disagree with that sort of thing. Like I’d say I ‘partially agree’, I ‘moderately agree’ and then I ‘strongly agree’.

[Interview 13: 192-203]

\*Interviewer: Do you know what you’d call them? 1, 2 and 3? If you could call them?

[...]

\*Respondent 13: ...maybe ‘neutral’ is like you say ‘don’t really agree, nor disagree’...1 perhaps is ‘mildly agree’, like ‘just about’ or ‘a little more than usual’...2 is ‘fairly agree’ – oh that doesn’t sound – ok ‘fairly certain’ that I agree with it...3 a little more, but 4 is really, ‘there is absolutely no doubt in my mind’.

Interviewee 11 also had a personally meaningful way of looking at agreement/disagreement which was grounded in the discipline he was studying, namely environmental sciences:

“\*Interviewer: What about the rest of the scale, so let’s see...you chose to put +3 and -3. How did you come to do that? What went through your mind?

\*Respondent 11: Oh well um, bipolar analysis is part of my course and is always +3 and -3 so I –

\*Interviewer: Did you say bipolar analysis?

\*Respondent 11: Yeh, I suppose that’s what this is...bipolar analysis.

\*Interviewer: Right ok, that’s really good. So it’s something you felt that your course influenced?

\*Respondent 11: Yeh. And I think that if there were any more states on that, there would be too many. But if there were any less, there’d be too few. So it’s the right amount.

\*Interviewer: Yes, and how does that... – explain a little bit about what bipolar analysis is on your course.

\*Respondent 11: For example if you were assessing pollution or something, you would say +3 if it were really clean and -3 would be totally polluted, and then there’s neutral yeh.”

[Interview 11: 115-141]

This was an interesting finding, given that his study discipline had exposed him to other types of rating-scales and it influenced the way he gradated his opinions about other concepts. As such, it was decided that in the quantitative phase of this research, that students from various disciplines be compared on their use of the IRSP.

This raised an additional point for consideration; the effect of a respondent's familiarity with fixed rating-scales, on their IRS. For example, those who might be very familiar with seeing seven-point Likert-type rating-scales may be predisposed to defining an IRS of -3←0→3. The drawback here could be that they might not be defining their true ideal rating-scale. However, given the IRSP instructions prompt respondents to introspect when anchoring their IRS for personally meaningful scenarios, it is likely that this would be avoided. Additionally, should respondents be provided with an option to re-modify their IRS, this could circumvent this issue.

Even more worthy of note was the implication that the IRSP, as a measurement method, could potentially augment respondents' *involvement* in the survey process. Increased involvement would result in respondents paying more attention to the questions being asked of them, and in turn they would think more carefully about their responses. Interviewee 12 said "you know when you do questionnaires and things, you don't really think about it" [12: 45-47] when referring to the amount of thought that goes into her responses to surveys in general. At a later stage of the interview she was asked a question, unrelated to the issue of involvement, yet her response catalysed a new direction.

\*Interviewer: ...Was there anything you found difficult about the exercise?

\*Respondent 12: No it was fine. It was actually just, sort of, thinking about things more. That was the only, sort of, challenge...actually, I don't know, like normally you wouldn't really care how many there was on the scale, because

you'd always answer questions based around those constraints that you've been given. But when you can actually decide what those constraints are, that makes it more difficult, it means you have to think about it more.

\*Interviewer: Mmhmm. Do you think that because you have to think about it more, as you just said, that it might make you think about your answers more?

\*Respondent 12: Yeah it does. I felt more aware about what I was saying, looking at those questions and looking at the scale that *I* had made.

\*Interviewer: Is that because you designed the scale?

\*Respondent 12: Yeah.”

[Interview 12: 353-372]

Initially she started to explain that the IRSP is more challenging than a typical questionnaire experience. Whilst this is potentially a drawback, the benefit is that she was clearly more aware of her opinions and able to express them in a more personally meaningful way (notice the stress on the “I” in the extract). Later, she went on to explain that because she was able to design her own rating-scale, she felt more involved in the questionnaire process and paid more attention to her responses than she would do in typical surveys.

The implication is that, if others feel the same way, this could improve the quality of the responses obtained from survey participants, both in terms of the *meaningfulness* of the answers and in terms of the amount of *attention* they give each question (i.e. increased involvement would result in more carefully considered responses as evidenced by the above interviewee).

#### *Greenleaf item nine*

Interviewee 12 made a very interesting observation with Greenleaf's ninth item; “TV advertisements place too much emphasis on sex.” Whilst in the process of rating the items, she paused to ask,

“\*Respondent 12: So, do you know here where you say ‘sex’, is it as in ‘gender’ or as in ‘the act’?”

[Interview 12: 123-125]

She was informed that the latter is probably what was meant. This was a very interesting query given the item could be interpreted either way. However, Greenleaf (1992b) in his fine-tuning of the large bank of items, reducing them down to an uncorrelated sixteen, will have determined that this item was independent enough from all other items, regardless of how it was being interpreted. So it could be argued that the item is still fulfilling its purpose as being uncorrelated to all other items, despite the potential for the word ‘sex’ to be interpreted in two ways.

#### *Clarity of Instructions*

One issue was raised with regard to the clarity of the IRSP instructions. This was by Interviewee 11 (the one who had misinterpreted the verbal anchoring instructions, and who had been observed skim reading). He stated, “I think it was generally clear, but like I said bits could be put in bold to make it clearer,” [Interview 11: 155-157]. This was a useful suggestion. Skim-reading is inevitable with some respondents, and so the use of underlining text and emphasising certain words in bold, might encourage respondents who skim-read to notice key elements within the instructions. It was decided the computerised version of the IRSP would have these key phrases highlighted. There were no other interviewees who explicitly suggested any improvements, when asked to consider this.

The general feedback was that interviewees found the IRSP easy to follow and execute, and all interviewees, bar Interviewee 11, indicated that they only needed to read the instructions once.

#### 4.2.8.2 *Key modification*

##### *The inclusion of a bar chart in the electronic IRSP*

It was clear that being able to practice using their IRSs on Greenleaf's items was very useful to respondents. It was decided that this feature should be incorporated into the electronic IRSP. Additionally, respondents were provided with the option to modify their IRS after using it to rate Greenleaf's statements. This check would make sure that respondents were happy with their IRS before proceeding to the main survey. It would also be interesting to see what portion of respondents needed to modify it, in that if no respondents chose to do so, this feature could be removed in future.

In order to help respondents make this decision, it would be useful if they could *see* how they were using their IRS. It was decided that the simplest way to *show* them would be to present them with a bar chart illustrating how often they used each of their intervals when responding to Greenleaf's statements. The accompanying instructions would need to encourage respondents to reflect upon the meaningfulness of their IRS, using the bar chart to help them. This planned IRSP modification is best illustrated by Figure 4. 18. This shows a screenshot, from the original software specification, of what was planned for the visual appearance of this facility. The example demonstrated what would be presented to a respondent who chose an IRS of  $-3 \leftarrow 0 \rightarrow 3$  with verbal anchors 'totally agree' and 'totally disagree', and who responded to Greenleaf's statements in the manner shown by the chart.

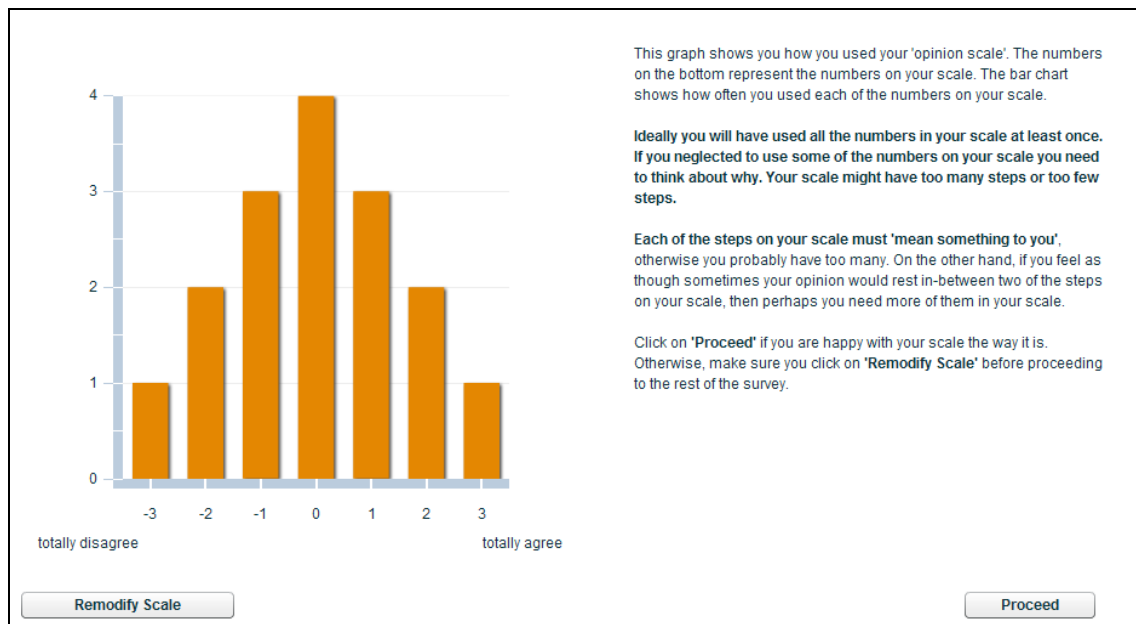


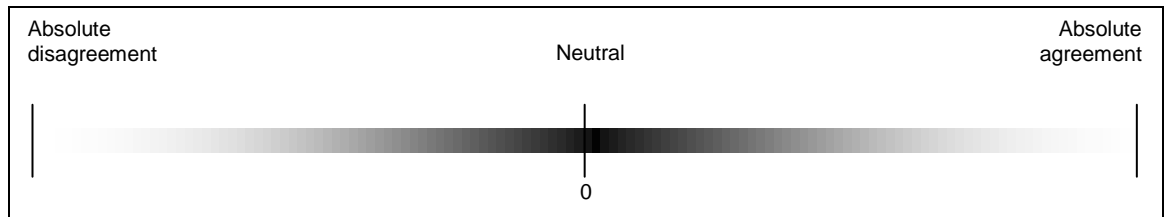
Figure 4. 18 IRSP bar chart visual aid

It was planned that the IRSP software be capable of allowing the researcher to switch on/off the use of the bar chart visual aid, so that if it proved to be unhelpful, the IRSP could also be executed without it.

#### 4.2.8.3 Potential improvement

##### IRSPv2

In order to best explain how this second version of the IRSP came about, imagine the agreement/disagreement continuum as a bipolar spectrum from white to black on each pole. For the purpose of this example, an individual's extreme position at both ends will be referred to as their *absolute disagreement* and *absolute agreement*. As shown in Figure 4. 19, *absolute disagreement* and *absolute agreement* are represented by the colour *white* (pure white). The neutral position represents the absence of *white*, and is therefore represented by *black*.



**Figure 4. 19 Agreement/Disagreement continuum represented as a spectrum of shades**

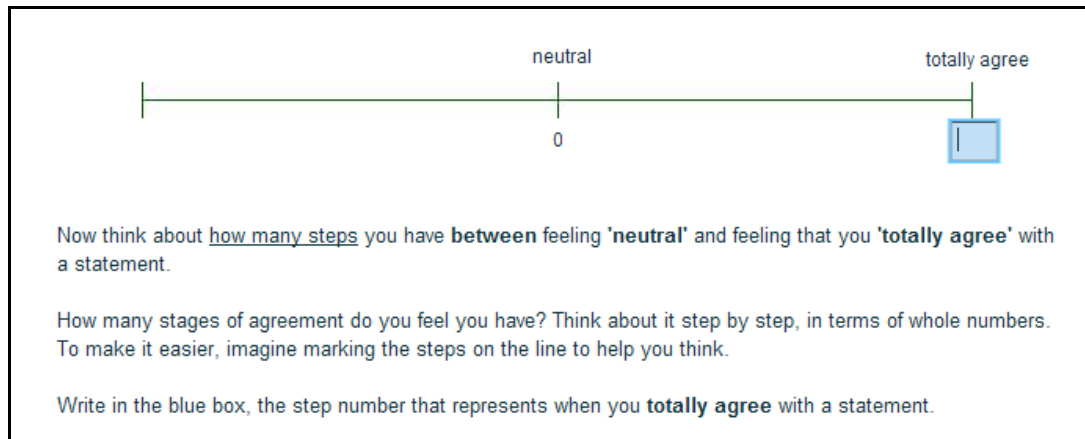
As this stage, the IRSP had respondents attach verbal anchors to both extremes; the *whites*. It also required them to assign numerical values that represent these *absolute* positions; one for the left-most *white* and one for the right-most *white*. Whilst it is clear from the interviews that many respondents think about the number of meaningful steps they have between *neutral* and their *absolute* positions (i.e. their *shades of grey*), some do not. Those who have elected to use  $\pm 10$  (a 21-point scale) have not been able to evidence that their eighteen *shades of grey* are all individually meaningful. When considering how to reduce respondents' attraction to anchor a *white* with a  $\pm 10$ , it was clear that the numerical anchoring instructions could be modified by shifting their focus from the anchoring of the *whites*, to the anchoring of the *shades of grey*. It was hoped that this would better ensure that all respondents give due attention to their *shades of grey*.

The numerical instruction for the anchoring of the absolute agreement position was worded as follows (D in Figure 4. 15):

Now think about how many steps you have between feeling 'neutral' and feeling (a) towards a statement. How many stages of agreement do you feel you have? Think about it step by step, in terms of whole numbers. To make it easier, you can mark the steps on the line to help you think. Write the number that corresponds with feeling (a), in box (b).



Whilst the instruction explicitly refers to the *shades of grey* (e.g. steps you have *between* feeling...), the numerical anchor respondents eventually assign represents *white* (i.e. the number that corresponds with feeling (a)). It was decided that the above IRSP instruction, as is, would be electronically represented in the way shown in Figure 4. 20.



neutral totally agree

0

Now think about how many steps you have **between** feeling 'neutral' and feeling that you 'totally agree' with a statement.

How many stages of agreement do you feel you have? Think about it step by step, in terms of whole numbers. To make it easier, imagine marking the steps on the line to help you think.

Write in the blue box, the step number that represents when you **totally agree** with a statement.

**Figure 4. 20 IRSP Numerical anchoring instruction represented electronically.<sup>4</sup>**

In order to see whether it would be more effective to have respondents focus their attention on anchoring the *shades of grey*, a second version of the IRSP would be tested (IRSPv2) alongside the original version (IRSPv1), in the next stage of qualitative research. Figure 4. 21 illustrates the modifications that were made to the instruction wording and the visual aid, forming the IRSPv2.

<sup>4</sup> For the purpose of this example, it is assumed that a respondent has already verbally anchored the agreement pole; 'totally agree'.



The same process would then be repeated for the disagreement pole. It was hoped that this might help respondents to focus more on the *meaningfulness* of each interval, and would probably result in more meaningful IRSs.

#### 4.2.9 Insights into the Conceptualisation of Agreement and Disagreement

When probing interviewees' reasons for choosing their verbal and numerical anchors, a deeper understanding was gained with regard to the way in which they conceptualised *agreement* and *disagreement*. It became clear that respondents fell into one of two camps; those that view *agreement* and *disagreement* as being 'mirror opposites' (bipolar continuum), and those that view them as being two 'different things' (two unipolar continuums). This finding had implications for the *usefulness* of the IRSP.

##### 4.2.9.1 Bipolar

When exploring the reasons for respondents' choice of verbal and numerical anchors, many indicated the need for a 'balance' and that they view *agreeing* and *disagreeing* as being opposites:

“\*Respondent 13: Yeah. I think you either – yeah, I think for a scale to work you need...uh, I can't think of the word...like and equal balance between the two...

\*Interviewer: Like an equilibrium?

\*Respondent 13: Thank you. [...] Yes.”

[Interview 13: 217-223]

Interviewees 1, 2, 4, 5, 6, 7, 9, 11 and 13, all regarded *agreement* and *disagreement* in a bipolar fashion. All, except Interviewee 4, chose the same absolute numerical value for both sides of their rating-scales (e.g.  $-4 \leftarrow 0 \rightarrow 4$ ,  $-10 \leftarrow 0 \rightarrow 10$ ). However, Interviewee 4's reasons for choosing -8 (and not -10) were outlined earlier and appeared to be linked to his inability to think of a meaningful verbal anchor. Additionally, all, except

Interviewees 1 and 4, chose the exact same verbal label for both sides of their rating-scale. One respondent (Interviewee 1) was under the impression that she *had* to choose different verbal anchors. Table 4. 7 illustrates some of the views held.

**Table 4. 7 Bipolar view of agreeing and disagreeing: Extracts from interviews**

<b>Interview: Lines</b>	<b>Discussion</b>
[5: 175-188]	<p>*Respondent 5 Yeh, if you totally agree with something, the ‘total’ is the most you could agree with it. The most you could disagree with something would be the same word.</p> <p>*Interviewer Right, umm...and -10 again, I guess because –</p> <p>*Respondent 5 It’s a complete mirror image, yeh.</p>
[6: 98-108]	<p>*Interviewer ...And you put the same one in next to ‘disagree’ –</p> <p>*Respondent 6 Yeh because that’s the opposite, isn’t it.</p> <p>*Interviewer Right, so you see agree and disagree as mirror opposites?</p> <p>*Respondent 6 Yep [...] Hence the numbers.</p>
[7: 257-269]	<p>*Interviewer Umm...do you feel that agreeing with something and disagreeing with something are mirror opposites? Or do you feel they are two different things?</p> <p>*Respondent 7 Mirror opposites.</p> <p>*Interviewer That’s why you’ve used the same adverbs?</p> <p>*Respondent 7 Yep.</p> <p>*Interviewer And that’s why you’ve also used the same numbers, positive and negative?</p> <p>*Respondent 7 Yep, exactly.</p>
[11: 272-297]	<p>*Interviewer You don’t think that they are two completely different things, you think they are like mirror images, like opposites on the same spectrum?</p> <p>*Respondent 11 I suppose you could kind of look at it a bit like love and hate...they’re kind of focused on something...but obviously they’re different...but I think more with agree and disagree they are opposites...yes, especially if you’re going for a scale like this, it’s either one or the other.</p>
[13: 213-238]	<p>*Respondent 13 I need a balance.</p> <p>*Interviewer You need a balance.</p> <p>*Respondent 13 Yeah. I think you either – yeah, I think for a scale to work you need...uh, I can’t think of the word...like an equal balance between the two...</p> <p>*Interviewer Like an equilibrium?</p> <p>*Respondent 13 Thank you. [...]</p> <p>*Interviewer [laughs] Ok so you feel that they’re mirror opposites of one another?</p> <p>*Respondent 13 Yes.</p> <p>*Interviewer You do. You don’t feel they’re two different things?</p> <p>*Respondent 13 Umm...no, no not at all.</p>

Naturally, respondents' conceptual regard for *agreement* and *disagreement* influences their choices when defining a rating-scale. The extracts seemed to indicate that those who regard *agreeing* and *disagreeing* in a bipolar fashion, are likely to define rating-scales that are numerically and verbally symmetrical.

4.2.9.2 *Unipolar*

Interviewees 3 and 12 indicated a slightly different way of looking at *agreement* and *disagreement*. Table 4. 8 contains some example extracts that demonstrate how *they* regard *agreeing* and *disagreeing*.

**Table 4. 8 Unipolar view of *agreeing* and *disagreeing*: Extracts from interviews**

Interview: Lines	Discussion
[3: 233-258]	<p>*Respondent 3 So I kind of toyed with the idea of making it the same and then for some reason I chose a different word.</p> <p>*Interviewer Can you figure out perhaps why you did that?</p> <p>*Respondent 3 Umm...I don't know...I'm seeing agree and disagree as different types of things.</p> <p>*Interviewer Ok. Do you see agree and disagree as opposites on a spectrum or do you see them as different concepts, for example...um...do you see them as mirror images agreeing and disagreeing?</p> <p>*Respondent 3 No. Not really actually, when I think about it. Not really, no. Maybe that's why. Subconsciously I'm treating them as different situations.</p> <p>*Interviewer And that's why you feel you had to put 'totally', a different word.</p> <p>*Respondent 3 Yes.</p>
[12: 87-99]	<p>*Respondent 12 I always feel they're sort of two different things...I don't know why...I always think they're like two separate things the way of like, the questions on like that have been worded separately. They are two different things, and it is like about feeling...like the feeling of agreeing and the feeling of not agreeing, so they are different 'cos they are different emotions.</p> <p>*Interviewer Right, that's really interesting...so you think they are different emotions? So you don't see them as mirror opposites?</p> <p>*Respondent 12 No I don't think anything is really that, like, clearly defined or anything...like in life...whether you agree or disagree with something is not as easy as a yes or no.</p>

Interviewee 3 had chosen the verbal anchors ‘completely agree’ and ‘totally disagree’ with a balanced numerical anchoring of  $-10 \leftarrow 0 \rightarrow 10$ . She did indicate, after rating Greenleaf’s statements, that her rating-scale had too many intervals and that she would have liked to change it to  $-5 \leftarrow 0 \rightarrow 5$  (thus she would still have maintained the numerical symmetry). She indicated quite clearly that the reasons for choosing different adverbs stemmed from her view that she sees *agreeing* and *disagreeing* as two different experiences. Interestingly however, she still desired numerical symmetry.

Interviewee 12 had also chosen differing verbal anchors; ‘definitely agree’ and ‘completely disagree’. Her numerical anchoring was also balanced as she chose a rating-scale of  $-5 \leftarrow 0 \rightarrow 5$ . Similar to Interviewee 3, although she chose differing verbal anchors she opted for balanced numerical anchors, and when asked *why* she asserted the following:

“\*Interviewer: Yeah, ok. You obviously put 5 and -5, um, any particular reason why you didn’t make one bigger than the other?

\*Respondent 12: That’s quite strange because I said before I didn’t see them [agreeing and disagreeing] as being the same, and I don’t...but at the same time there needs to be this balance of how you define them – [laughs]”

[Interview 12 : 109 - 115 ]

Interestingly it would seem that despite viewing both *agreeing* and *disagreeing* in a unipolar fashion (i.e. two unique poles), she still preferred numerical symmetry. Whilst it appears that most respondents would prefer to have a numerically balanced rating-scale, an advantage of the IRSP is that it affords respondents the option to choose otherwise.

#### 4.2.9.3 Unipolarity, bipolarity and the IRSP

The interviews suggest that whilst most respondents have a bipolar view of agreeing/disagreeing, there are some that regard them as unipolar. Typically the researcher would need to impose his/her numeric choices, and thus conceptualisation, on the respondents through the selection of a fixed rating-scale. However, the IRSP allows respondents to define their *own* rating-scale verbally and numerically, in a way that reflects *their* conceptualisation of agreement and disagreement. Even though the IRSP always has *neutral* anchored at 0, this will suit both those with a unipolar and a bipolar conceptualisation. A respondent with a bipolar view of *agreement/disagreement* is likely to choose the same adverbs for either side of the rating-scale, and is also likely to choose a numerically balanced IRS (i.e. the same number of intervals either side of neutral). However, should a respondent see both *agreeing* and *disagreeing* as *different* feelings (rather than as opposites), they can have a differing number of meaningful intervals for each. The IRSP can accommodate this. For example, should they feel that they have three meaningful levels of *disagreeing* and four meaningful levels for *agreeing*, they could define an imbalanced IRS of  $-3 \leftarrow 0 \rightarrow 4$ . In this scenario the neutral point still works well, as it represents the absence of either (as is done with bipolar continuums) whilst still allowing a form of unipolar conceptualisation. This appears to be a clear advantage that the IRSP has over researcher-defined fixed rating-scales.

### 4.3 Stage 2: The IRSP from paper to software

The next stage of development was to design a software specification so that IRSP survey software could be undertaken. This section details the main stages involved in transforming the IRSP into a working computer program.



### 4.3.1 Development of the IRSP Software

A specification was designed, and a suitable programmer was contracted to execute it. The original specification can be seen in Appendix E. The specification detailed the following key requirements:

- Users needed to be able to individually define a rating-scale (both the number of intervals and the verbal endpoints), in the manner described in the specification;
- Users needed to be presented with a visual representation of their IRS and use it to answer a set of questions, in the manner described in the specification;
- The data captured from the responses needed to be stored into a database, structured in the manner described in the specification;
- The researcher needed to be able to modify the questions included in the survey.

The programmer provided a satisfactory time-estimate on the work required to fulfil the specification and also provided other documents for the researcher to inspect such as the planned database structure (e.g. see Appendix F for a relational diagram). The programmer created basic IRSP software, based on the original specification. It was tested for errors. Errors were highlighted, with instructions for the corrections detailed in a report. See Appendix G for an example of one of the error reports. These reports would often be supported with a discussion over the best course-of-action to rectify the errors. Upon inspection of the prototype version of the IRSP program, it was decided that it could be developed further. This was done so that it could be used in a multi-group experimental design for quantitative testing of the IRSP, which would enable it to be compared to typical fixed rating-scales. This meant that the software would need to be able to accommodate both the use of Likert-type rating-scales and the IRSP in survey design. Also, the potential to create IRSP software that could be used beyond the scope of this PhD project was useful, for future research.

### 4.3.2 The Finished IRSP Software

The IRSP Software consisted of a back-end ‘survey management system’ where the researcher has the facility to create online surveys using the IRSP measurement tool, as well as typical Likert formats, and a front-end ‘survey platform’ where respondents can complete the online surveys created. The IRSP software end-product was more than satisfactory and met all the specification requirements for the project. The researcher purchased the domain name [www.phdsurvey.co.uk](http://www.phdsurvey.co.uk) for the web hosting of the survey management system and databases. A detailed explanation of the features, with illustrations, can be found in Appendix H.

On visiting the following page <http://phdsurvey.co.uk/SurveyAdmin/index.html>, the researcher is presented with an entry-restriction requiring a username and password. This facility allows the researcher to modify/create surveys and monitor databases on the internet. The username and password protects unauthorised access.

The following list summarises some of the key features of the software detailed in Appendix H.

- Researcher-flexibility over survey design.
  - Create/modify questions and manipulate ordering.
  - Choose from a menu of demographic items for inclusion.
  - Assign one of several measurement methods to question pages; Likert-5, Likert-6, Likert-7, IRSP or IRSP2.
  - Create/modify instruction pages, manipulating features such a font, colour, font size.

- Weave hyperlinks into text within the survey instruction pages.
- Assign a unique html address for every survey created.
- Copy entire surveys, saving the need to re-create an identical survey from scratch for a different sample population.
- When survey questions are assigned the IRSP or IRSPv2 measurement method, respondents are able to individually define a rating-scale for both the number of intervals and the verbal endpoints;
  - Respondents are presented with a clear visual of their IRS and can use it to answer the survey questions, in the manner described by the specification;
- The respondent-data captured and stored in the databases are displayed clearly and can be copied into Excel or SPSS easily.

#### **4.4 Stage 3: Further development of the IRSP**

Insights from Stage 1 highlighted two possible variations of the IRSP. Stage 3 was necessary to further consider these two emerging variations of the IRSP, IRSPv1 and IRSPv2, and examine respondents' reactions to them. It was also necessary to check both the user-friendliness of the IRSP Survey Software and the accuracy of the data captured. The overall objective of this stage was to determine whether the IRSP survey software was performing properly.

##### **4.4.1 Key Questions**

The purpose of Stage 3 was to address the following key questions:

- How well does the survey software perform?
- How does the IRSPv1 compare with the IRSPv2?

- Are both easy to execute?
- Does one, of either IRSPv1 or IRSPv2, encourage respondents to produce more meaningful IRSs?
- Do any parts of the instructions need further modification?
- Are the visual aids appropriate and helpful to respondents?
- Do the respondents fully understand all the questions?
  - Demographic items
  - Greenleaf's sixteen items
  - Other chosen scale measures

The answers to these questions: enabled further improvements to the IRSP software; helped towards determining which one of the two IRSP methods to test in the quantitative phase; and alerted the researcher to items that might be problematic.

#### **4.4.2 Psychological Measures**

In the Literature Review and Methodology chapters, the reasoning behind the decision to include particular measures of personal traits (in addition to demographics) was discussed. It was key that the measures chosen be already validated in the literature. In this way, the results could then be trusted to reflect the personality construct under examination. In addition, the uses this would have for a multi-group experimental design, in the main quantitative phase, were obvious. Some respondents could answer a version of the online survey using the rating-scales by which the psychological measures were already validated, and these results could be compared to those obtained when respondents' personality measures were captured using the IRSP. These measures

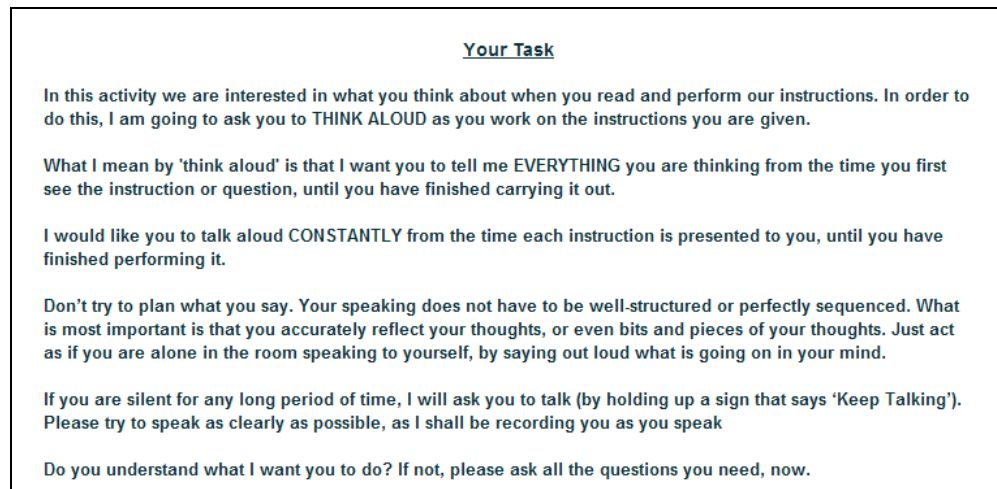
were included in the computer-administered version of the IRSP used in the interviews within Stage 3.

#### **4.4.3 Method**

##### *4.4.3.1 Concurrent Verbal Protocol-Retrospective Debrief (CVP-RD) Interviews*

In-depth interviews were conducted with sixteen respondents, using two forms of verbal report with each; concurrent verbal protocols (CVP) followed by retrospective debriefings (RD) (Taylor and Dionne, 2000, Bolton, 1993). “CVP are verbatim records of a problem solver thinking aloud while solving a problem. RD are the problem solver’s account of how a problem was solved, reported following the problem-solving activity,” (Taylor and Dionne, 2000: 413). Interviewees were asked to complete *part* of the online survey whilst ‘thinking aloud’ (CVP) and were subsequently interviewed about their experiences with the exercise, upon its completion (RD). These concurrent verbal protocol-retrospective debrief (CVP-RD) interviews provided the opportunity to test the respondents’ understanding of the scale items and the items’ suitability for inclusion in a larger survey.

After reading the project’s ethical code of conduct and typing in their personal details, interviewees were presented with the CVP instruction, shown in Figure 4. 23. To see the entire survey that was presented, please refer to the screen shots in Appendix I.



**Figure 4. 23 CVP Instructions to CVP-RD Interviewees**

Interviewees only had to 'think aloud' during part of the survey, specifically, whilst carrying out the instructions to define their IRS, rating Greenleaf's sixteen items, considering the information presented to them on the graph page, and if selected, whilst re-modifying their IRS. Interviewees were presented with instructions telling them to stop thinking aloud before proceeding to rate the fifty-six psychometric items. However, they were instructed to speak up if they came across any words or statements that were confusing, ambiguous, or that they did not understand.

In order to make respondents feel at ease and to encourage them to act as naturally as possible, the researcher stood behind them and to one side. This meant that respondents' verbal protocols could not be directed at the interviewer in a conversation-like manner. Standing to one side also avoided making the respondent self-conscious. The researcher stood far enough away to give the respondent space but close enough to observe the respondent's actions on each screen. The researcher spoke rarely during respondents' CVPs, only interjecting to hold up a 'Keep Talking' sign if fifteen or so seconds had passed in silence. The researcher took notes whilst observing the CVPs and frequently referred back to them during the RDs when probing themes.

#### 4.4.3.2 *CVP-RD Interview Setting*

The CVP-RD interviews took place in either an empty class room at the School of Management site reserved especially for the interviews, or at the Atrium building (a public space) in a specially reserved cubicle for privacy. In both of these settings, all interviewees appeared to be at ease during the exercise.

#### 4.4.3.3 *CVP-RD Interview Sample*

Sample size was not planned in advance; interviews would continue as long as they could significantly contribute to further developments of the IRSP software. It was important to continue to use analysis-driven purposive sampling and include students across varying stages of study; further development (to both versions of the IRSP) was aided by the experiences of different *types* of student. Additionally, it was deemed prudent to include some students whose first language was not English. Given that foreign students represent a portion of the student body, it was considered appropriate to see if they had any problems carrying out the IRSP.

As can be seen in Figure 4. 24, Figure 4. 25, and Figure 4. 26, a satisfactory range was achieved across age, gender and language when testing both IRSPv1 and IRSPv2.

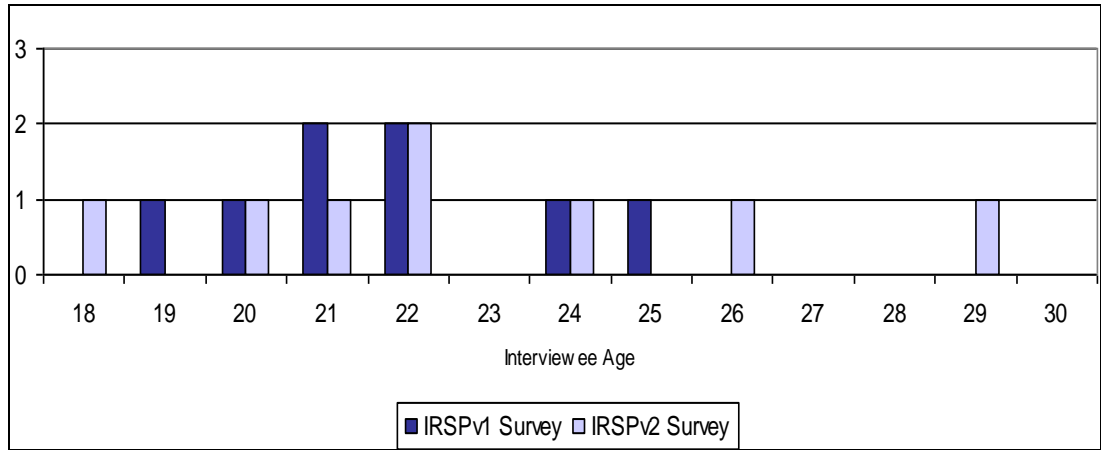


Figure 4. 24 CVP-RD Interviews: Age Spread

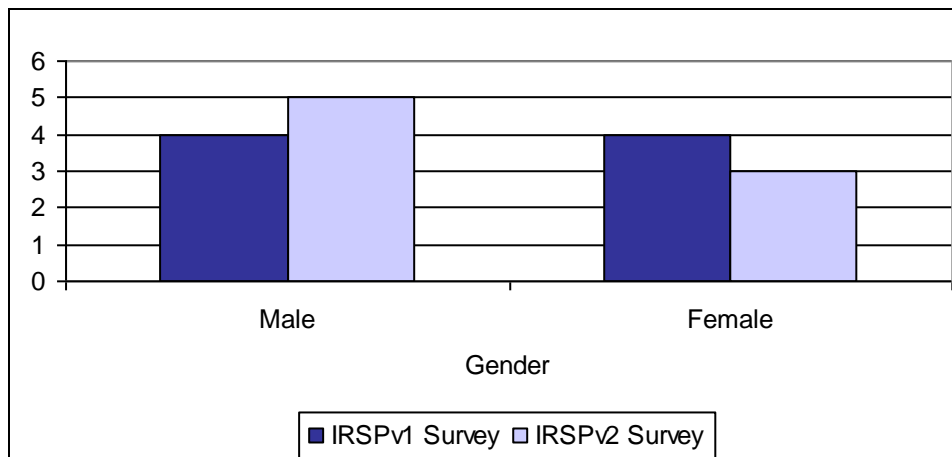


Figure 4. 25 CVP-RD Interviews: Gender Spread

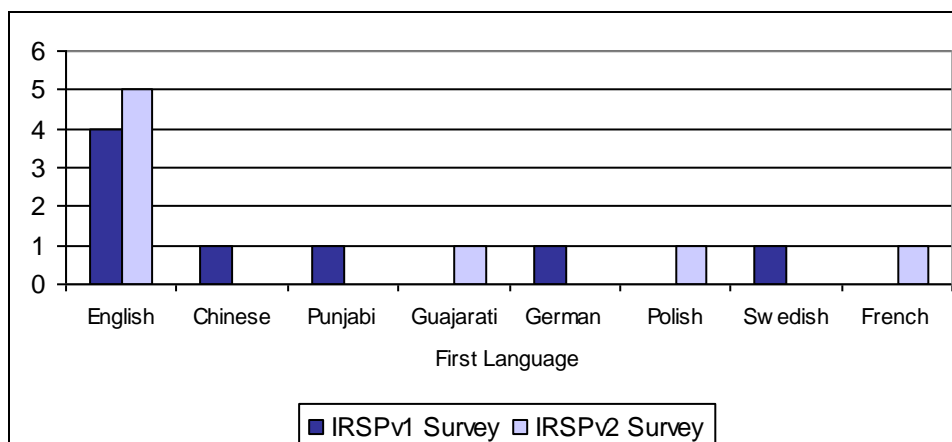


Figure 4. 26 CVP-RD Interviews: First Language Spread



#### **4.4.4 Analysis of Stage 3**

All sixteen protocol-debrief interviews were recorded using an unobtrusive digital voice recorder; permission was granted in all cases. Each interview lasted on average forty-five minutes.

It was not considered necessary to transcribe these interviews, as was done with Stage 1, as this stage was for tweaking and adjusting the method, rather than exploring issues in depth. As such, a closer inspection of the semantics was not needed in order to achieve the objectives of Stage 3. Because interviewees in this stage were thinking aloud throughout the entire process, this permitted the researcher to concurrently make notes about where a modification was likely to be needed. Interviewer-notes included operational observations (relating to the software) and behavioural observations (about the interviewees), which helped identify the likely modifications required. The researcher noted these likely modifications during the concurrent verbal protocol (CVP) part of the interview. This permitted the quick and efficient modification of the electronic IRSP to be carried out before the next round of protocol-debrief interviews. The observations made during the CVP also enabled the researcher to note down additional questions to raise during the interviewee's retrospective debrief (RD). Afterwards, the digital recordings were played back in order to ensure that the key findings had indeed been noted down, before proceeding. The live interview-notes were scanned and are included in the CD attached to this thesis.

#### **4.4.5 Findings**

##### *4.4.5.1 IRSP Survey Software performance*

On inspection of the CVP-RD notes, it was clear that small modifications to the survey software were necessary. These included: aesthetic changes to the way some demographic items were presented on screen; setting additional error prevention messages, for example, stopping respondents from accidentally typing a number into a verbal-anchor box; and fixing a few minor bugs on pages that loaded when IRSs were re-modified. Errors were listed for the programmer, and amendments were made to the software accordingly. Overall, the IRSP Survey Software performed well and data was captured accurately, as checks were carried out on the database of responses.

##### *4.4.5.2 IRSPv2 or IRSPv1?*

The quantitative data collected from this small sample of 16 respondents, together with the CVP-RD qualitative insights, appeared to suggest that the IRSPv2 was more effective than the IRSPv1 in having respondents produce more distinct response intervals. Interestingly, respondents using the IRSPv2 had a tendency to define IRSs with fewer categories than did those using the IRSPv1. The mean number of response categories for those that used the IRSPv1 was 11.38, with a standard deviation of 7.05. Whereas the mean for those that used the IRSPv2 was 8.88, with a standard deviation of 5.35. Figure 4. 27 helps to illustrate this point.

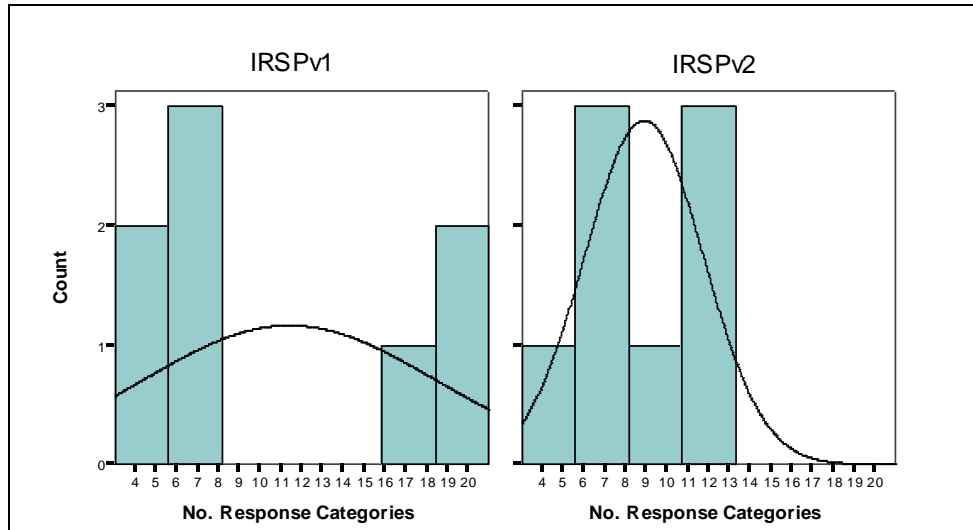


Figure 4. 27 Respondents' chosen number of categories on versions 1 and 2 of the IRSP

There were two people that did IRSPv1 who defined  $-10 \leftarrow 0 \rightarrow 10$  IRSs (i.e. 21-point rating-scales), and both did not wish to re-modify their IRSs even after seeing how rarely they used some of the intervals on their rating-scale. The spread of their responses are shown in Figure 4. 28 and Figure 4. 29.

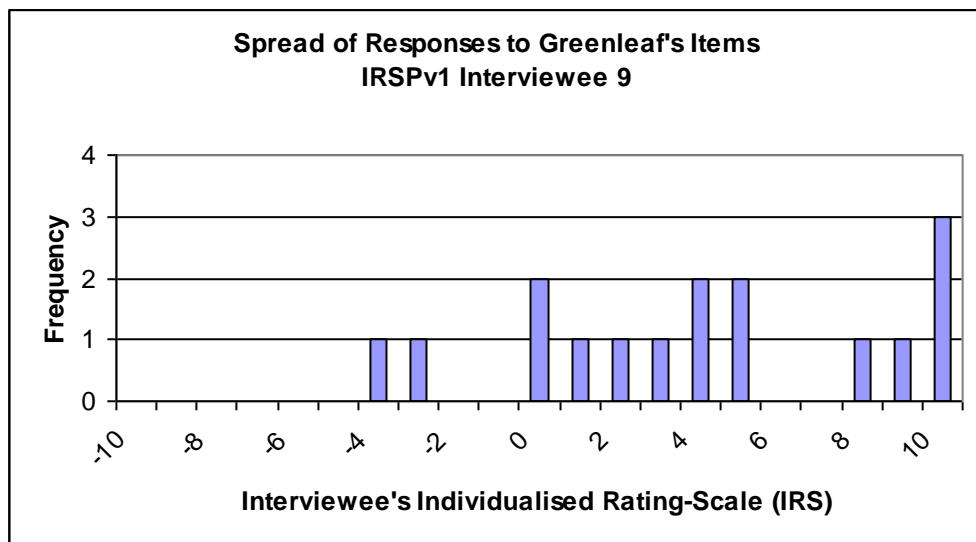
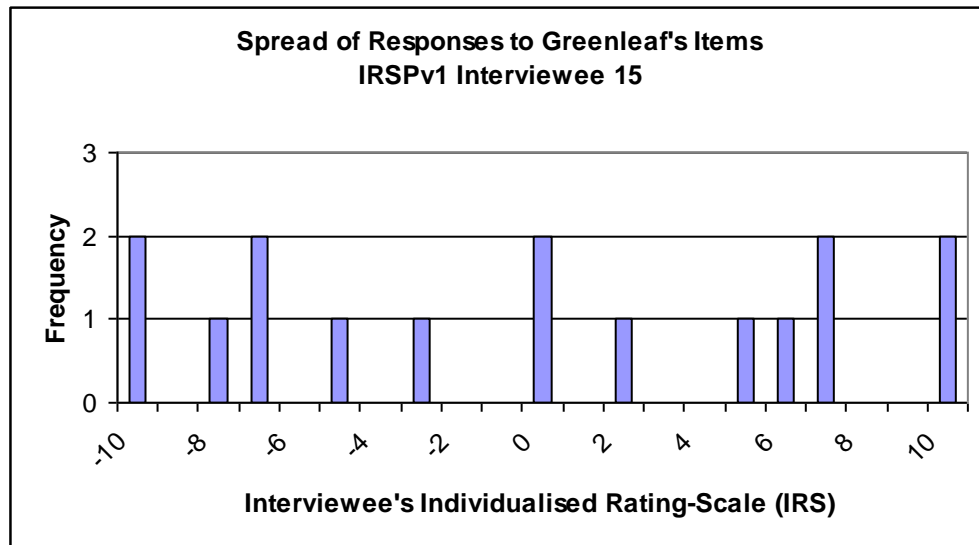


Figure 4. 28 CVP-RD Interviewee 9: Spread of responses to Greenleaf items



**Figure 4. 29 CVP-RD Interviewee 15: Spread of responses to Greenleaf items**

Both of these interviewees indicated that they felt that their rating-scales were adequate, and that there were not enough questions which caused them to feel the full extent of the varying levels of opinion represented by their rating-scale. When observing these interviewees during the numerical anchoring process, it was clear that they had chosen their numerical anchors without thinking about the number of steps they had *in-between* neutral and their absolute. They had immediately chosen a number which would quite nicely represent the endpoint, but without giving enough due consideration to the 'shades of grey'. This is where the IRSPv2 performed much better than the IRSPv1, given the focus of the numerical instructions effectively asked respondents to do just that; count the 'shades of grey'.

Respondents' attraction to the mystery  $\pm 10$  (in this automated fashion) with IRSPv1, echoed the findings from Stage 1. When asked why she chose '10', Interviewee 15 replied,

“I don’t know, things are always like from one to ten. [...] Cos it’s a round number, it’s easy to work with and my maths isn’t that good, so I just pick the one that’s easiest to divide by, or whatever, so I thought ten, and most scales are done from one to ten. [...] From when my sister used to do lots of surveys and stuff, hers were a lot by one to ten. So I’ve seen that a lot before. [...] And it’s a big round number, easy to use.”

Interviewees 9 and 15 both indicated that by choosing so many intervals, it made the task more demanding, and Interviewee 15 indicated that she would have chosen fewer intervals in hindsight. She also indicated that “quite a lot” of the intervals were not meaningful to her, and that she wished she had re-modified her IRS when given the opportunity. When asked why she did not, she eventually admitted it was because she did not really understand the graph, probably because she did not give it much attention.<sup>5</sup>

Some respondents explicitly stated that they found the IRSPv1 wording confusing when being asked to numerically anchor their IRSs. Interviewee 8 stated, “I don’t get that... “Write in the blue box the step number that represents when you strongly agree”. Does that mean – what does that mean?” He was informed that the instruction could not be clarified, because he had to complete it unaided and that there was no right or wrong answer. He re-read it and decided to assign ‘2’ in the box. In his RD he said that the instruction “wasn’t very clear in what it was asking you to do and also the way it was asking you to input the data in relation to the image of the scale on the screen...I think it confused me a little bit.” He then went on to say that he could tell immediately what he had to do next when he saw the visual aid with a blue box beneath ‘strongly agree’, but that the wording of the instruction confused him and made him re-think his initially correct assumption.

---

<sup>5</sup> Note: The graph page was subsequently made clearer in both versions of the IRSP.

On the other hand, the IRSPv2 instructions and visual aid for the numerical anchors seemed to be understood perfectly. Several interviewees even verbalised the verbal anchors they were mentally assigning to their ‘shades of grey’. For example, in his CVP, Interviewee 3 said the following, “so there’s neutral, and then it would go... ‘sort of’, ‘strongly’ then ‘totally’.” He therefore came to the conclusion he had two ‘shades of grey’, and typed two in the box. He repeated this process for disagreement, eventually ending up with a  $-3 \leftarrow 0 \rightarrow 3$  IRS. In his RD he indicated that he could have used an extra interval on either side (a  $\pm 4$  rating-scale) but that he did not because he could not think of a word that would slot in nicely among the other steps. More importantly, he added, because this extra step did not naturally come to mind immediately: “it [the interval] can’t be that important to me [...] life is like shades of grey, and if I didn’t manage to pick out a certain shade quickly, then maybe there is no need for me to have that shade.” A very insightful comment, and it is worth noting that he used the term ‘shades of grey’ unprompted (i.e. an *en vivo* phrase). Overall, the IRSPv2 appeared to be getting respondents to think far more meaningfully about their intervals than IRSPv1.

#### 4.4.5.3 *Summary of modifications to IRSP instructions*

The CVP-RD interviews resulted in several modifications to the IRSPv1 and IRSPv2 instructions, some occurring *between* interviews and some occurring at the end of the entire cycle. Table 4. 9 summarises the modifications made in the order they occurred.

**Table 4.9 CVP-RD Interviews: Resulting IRSP modifications made (in order).**

Versions corrected	Modifications made to the IRSP
Both IRSPv1 and IRSPv2	<ul style="list-style-type: none"> <li>• Corrected a spelling mistake on the word ‘constantly’.</li> <li>• Bolded and capitalised “word”, in the IRSPv1/v2 instructions for verbal anchors.</li> <li>• Added the phrase “make sure you” to the instructions on the ‘Option to Re-modify’ screen, to stress the importance of re-modifying their IRS if they are not completely happy with it.</li> <li>• “Do you feel as though you don’t #####* agree/disagree with someone but you agree/disagree with them to some extent?” was added to the numerical anchoring instructions.</li> <li>• Changed spelling error in Greenleaf item five from “food” to “foods”.</li> <li>• Improved the clarity of the wording in some of the demographic items.</li> <li>• Changed the paragraph order of the verbal anchoring instructions.</li> <li>• Increased the font size of the verbal anchoring instructions.</li> <li>• Modified the wording of the ‘neutral’ instructions.</li> <li>• Modified the instructions on the graph page to improve clarity; used titles, added labels to graph, and modified wording.</li> </ul>
IRSPv1 only	<ul style="list-style-type: none"> <li>• Emboldened and capitalised ‘step number’ in IRSPv1 instructions for numerical anchors.</li> </ul>

\* The survey software weaves whatever verbal anchors were chosen by the respondent into the instructions.

#### 4.4.5.4 Visual Aids

As expected, there were no problems reported with the rating-scale visual aids. In fact, it was clear in all cases that respondents were able to understand and execute the task as quickly as they did, *because* of the dynamic rating-scale visual aids. They would see their rating-scale *transform*, depending on their entries; this facilitated the entire IRSP process.

Feedback about the graph page and accompanying instructions<sup>6</sup> was very interesting. The majority of interviewees found the graph very useful, helping them to reflect on the meaningfulness of each interval on their IRS. Interviewee 1 was the only one that indicated that the instructions alone (without the graph) would have sufficed in terms of making him reflect on the meaningfulness of his IRS. However, in an earlier comment, he indicated that after seeing how he had failed to use some of his intervals (which he

<sup>6</sup> Where respondents, after completing Greenleaf’s sixteen items, are presented with a bar chart and a set of instructions asking them to reflect on the meaningfulness of their IRS: the x-axis is their IRS, and the y-axis shows the number of times they used each interval.

ascertained by *looking* at the *graph*), it made him question his IRS for a moment. After pondering whether it *was* meaningful, he came to the conclusion that it indeed *was*, and opted to keep it. Therefore, whilst he did not change his IRS, if the graph visual aid had not been provided, he may not have paused long enough to consider his IRS. For those that modified their IRSs, they all felt that the graph was pivotal in helping them to reflect on the meaningfulness of their intervals. Interviewee 2 said that the graph was “definitely helpful”. He said it made him realise that he would be better off making his rating-scale shorter, making the process less taxing. After shortening it, he found the rating process much easier, and his intervals were more meaningful to him. Of those that chose to modify their IRSs, none chose to *increase* the number of response categories. For those that felt as though they were satisfied with their IRS after rating Greenleaf’s statements, being presented with the graph page simply made them more convinced about the suitability of their IRS. For example, Interviewee 6 said that “seeing the graph simply confirmed what I already thought”, that all the intervals of her IRS were meaningful. Even for those who felt as though each of their intervals were meaningful, but had neglected to use one or more, if they were confident enough about their IRS, and despite the prompt suggesting they may need to shorten it, they opted to proceed with it as it was. Figure 4. 30 and Figure 4. 31 show the spread of responses for Interviewee 3 (both for Greenleaf and for the fifty-six main survey items).



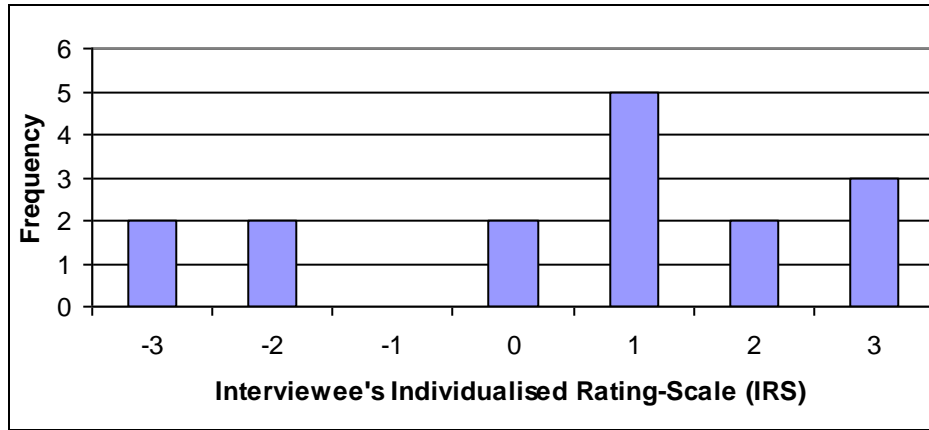


Figure 4. 30 CVP-RD Interviewee 3: Spread of responses for Greenleaf items

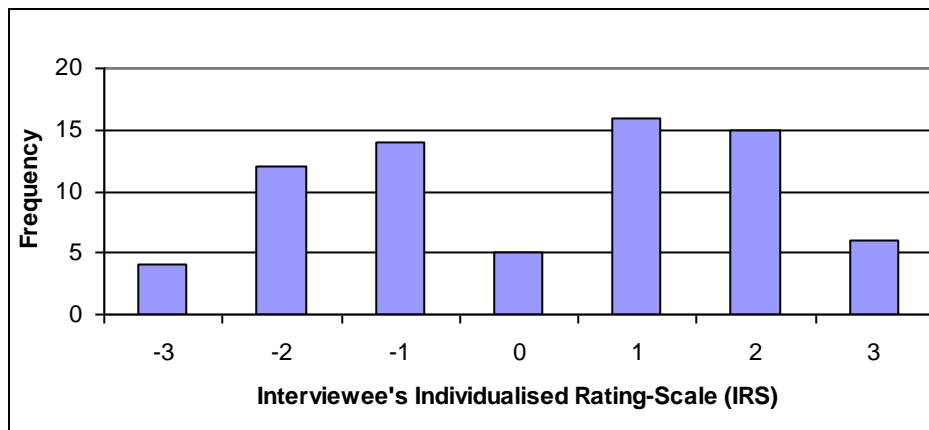


Figure 4. 31 CVP-RD Interviewee 3: Spread of responses for main survey items

Interviewee 3 could have potentially shortened his IRS if the graph page instructions had been more suggestive, but he was sure that -1 *was* meaningful to him and that he simply had not had the opportunity to use that position yet. In the responses to the main survey items, he clearly had no problems discriminating between all the intervals on his IRS, and indeed used the -1 position. It was clear that the graph page instructions were persuasive enough to entice those who were not completely happy with their IRSs, into re-modifying them. However, the graph page did not appear to dissuade those that were satisfied with their IRSs from continuing to use them, based on the observations of the CVP-RDs.

The graph page had a particularly interesting effect on respondents who appeared to adopt stylistic responding. After rating Greenleaf's items and being shown the graph page, stylistic responders seemed to become aware of the fact. One particularly interesting case was Interviewee 4. When prompted to reflect on the meaningfulness of his -3←0→3 IRS (after using it to rate Greenleaf's statements), he modified his numerical endpoints (feeling "more comfortable" with a numerically-imbalanced -2←0→3 IRS). He decided that he did not need quite as many levels of 'disagreement' as he did with 'agreement', and shortened his IRS accordingly. He stated in his retrospective debrief that, prior to seeing the graph, he had no idea that his responses were so skewed and that "I didn't realise I was such an agreeing person." When asked whether this realisation affected the way he completed the rest of the survey, he responded "I looked at the statements [psychological items] like I did with the first set [referring to Greenleaf's items] and just thought, "*do* I completely agree with this?" Um, I didn't want it to be as neutral as before. I wanted to be a bit more assertive with my answers." It was quite apparent that after being made aware of the fact that he had a tendency to acquiesce, he tried to be more honest about his *true* feelings. When comparing his spread of responses on Greenleaf's items using his -3←0→3 IRS (Figure 4. 32), to his spread of responses on the subsequent items using his modified -2←0→3 IRS (Figure 4. 33), it would seem that his tendency to acquiesce was reduced. However, the main survey encompassed a different set of items rendering a direct comparison impossible, so it did not necessarily ensure that acquiescence had not occurred, but it is a positive indicator nonetheless. On Greenleaf's items (for which one should have a uniform spread of responses, if response bias is not present), he did not rate a single item greater than -1 on his negative pole. However, the second graph shows a noticeable difference. The graphs, taken together

with his verbal reports, suggest that is reasonable to conclude that his tendency to acquiesce may have reduced as a result of the IRSP.

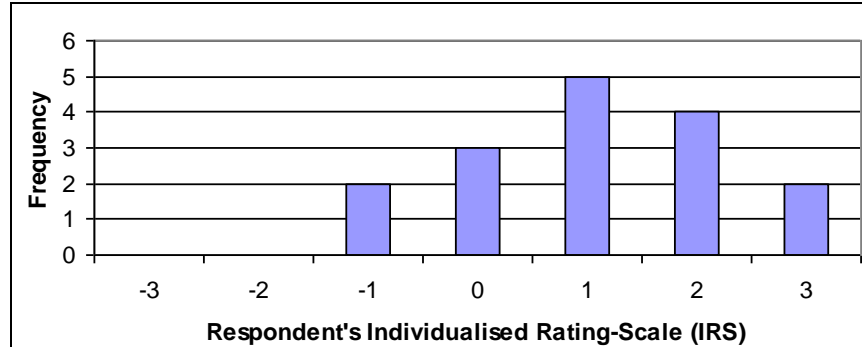


Figure 4. 32 IRSPv1 CVP-RD Interviewee 4: Spread of responses for Greenleaf's items.

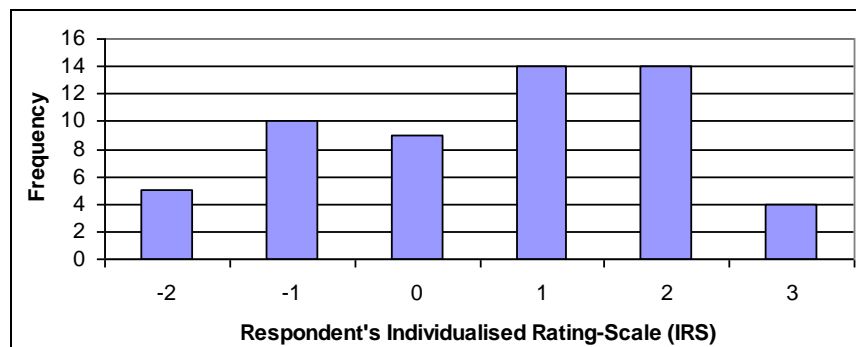


Figure 4. 33 IRSPv1 CVP-RD Interviewee 4: Spread of responses for main survey items.

4.4.5.5 *Problems reported with items*

Some interviewees did not understand certain words and others found that some statements were ambiguous. Table 4. 10 summarises these items that presented issues, the issues raised, and the first language of the interviewee (as this might have had some bearing on why some words may not have been understood).

Whilst some of the items flagged were identified only by interviewees whose first language was not English, three of those items were also flagged by native English

speakers; items CoSI\_17\_C, PNS\_10R and AO\_11. Additionally, there were items flagged by more than one non-native English speaker; AO\_14, BFI\_10, CoSI\_10P. These items were checked later for their factor loadings, when testing the measurement model in the quantitative phase.

**Table 4. 10 Protocol-debrief interviews: Problems with items.**

Interviewee.	First Language	Items flagged	Details
1	French	AO_14	Did not understand “subtle”.
2	Punjabi	BFI_10 AO_11 CoSI_10_P CoSI_17_C PNS_10R	Found “active imagination” ambiguous. Did not like the wording of AO_11. Did not understand “meticulously”. “I like to extend boundaries” was ambiguous. Did not understand “exhilaration” and guessed it might be a word for ‘chance’.
3	English	BFI_09 CoSI_17_C PNS_10R	Questioned whether BFI_09 means ‘showing one’s nerves externally’ or ‘being nervous inside’. I noticed him pausing for a long time on the ‘extend boundaries’ items (CoSI_17_C). Did not like the word ‘exhilaration’, and thought of it as ‘buzz’ or ‘excitable’.
9	English	AO_11	Found AO_11 confusing.
10	German	BFI_07 PNS_05R	“Fault with others” was ambiguous; he interpreted it in the context of ‘blame’ and not ‘criticism’. “Tedious”, he interpreted it as ‘boring’.
12	Swedish	BFI_02 BFI_08 BFI_10 AO_14 CoSI_10_P CoSI_17_C	She interpreted “trusting” as ‘trustworthy’. Did not understand “thorough”. “Active imagination”: Said it could be interpreted in two ways; one who is creative; one who over-analyses things. Did not understand “subtle”. Did not understand “meticulously”. Did not understand “I like to extend boundaries”.
14	English	CoSI_17_C	“I like to extend boundaries” was ambiguous.

NB Respondents 4-8, 11, 13, 15 and 16 did not report any problems with items.

#### 4.4.5.6 Greenleaf’s items and the IRSP

Allowing interviewees to practice using their IRS on Greenleaf’s sixteen uncorrelated items proved a valuable part of the process for several reasons;

- Interviewees could ascertain the ease-of-use of their IRS.
- Interviewees could reflect on their responses to the items, and consider whether their response categories were *distinctly* meaningful to them, before proceeding with the rest of the survey. This stage formed part of the facility which allowed them to ‘reduce’ or ‘increase’ the number of categories on their rating-scale.

- The CPVs showed that respondents who had a tendency to acquiesce, for example, became conscious of it when reflecting on their use of their IRS on Greenleaf's items. In some of the retrospective debriefs, these respondents said that after realising this, they subsequently tried to be more honest and accurate about their opinions when rating items in the main survey.

#### *4.4.5.7 Verbal anchoring and meaningfulness*

Generally, interviewees were easily able to anchor their own verbal endpoints on the agree/disagree continuum in both versions of the IRSP. There were only three who chose atypical verbal endpoints, shown in Table 4. 11 (Interviewees 5, 11 and 13). It is worth reiterating that the verbal-anchoring instructions were the same in both versions of the IRSP.

**Table 4. 11 Verbal endpoints chosen by respondents.**

Interv.	Age	Gender	First Language	Ethnicity	National Identity	IRSPv1	
						Verbal Endpoint (agree)	Verbal Endpoint (disagree)
1	22	Male	French	Other	French	totally	totally
4	24	Male	English	White	English	definitely	completely
6	22	Female	English	White	English	absolutely	Absolutely
8	21	Male	English	White	English	strongly	Strongly
9	25	Male	English	White	English	Fully	Fully
13	20	Female	Polish	White	Polish	minimally	slightly
15	21	Female	Guajarati	Asian - Indian	British	totally	highly
16	19	Female	English	Black - African	English	totally	completely

a

Interv.	Age	Gender	First Language	Ethnicity	National Identity	IRSPv2	
						Verbal Endpoint (agree)	Verbal Endpoint (disagree)
2	21	Male	Punjabi	Asian - Pakistani	British	totally	fully
3	24	Male	English	White	British	totally	totally
5	26	Male	Chinese	Asian - Chinese	British	ok	not ok
7	18	Female	English	White	English	totally	totally
10	22	Male	German	White	English German	Fully	fully
11	22	Male	English	White	English	Agree	Strongly
12	29	Female	Swedish	White	Swedish	definitely	totally
14	20	Female	English	White	English	definitely	completely

b

c

<sup>a</sup> This respondent indicated that she misunderstood the verbal anchoring instructions.

<sup>b</sup> This respondent indicated that he misunderstood the verbal anchoring instructions.

<sup>c</sup> This respondent indicated that he had misunderstood the first verbal anchoring instruction.

In their retrospective debriefs, Interviewees 5 and 13 felt that they had misunderstood the verbal anchoring instructions as a result of not having read the instructions properly and possibly the language barrier (given that their first languages are Polish and Chinese, respectively). Interviewee 11 apologised for his seemingly unusual “agree” verbal endpoint, explaining that he had rushed through the exercise and had not read through all the instructions. Nonetheless, the three interviewees indicated that when rating statements, they treated the endpoints on each side of the continuum as *the most they could possibly agree/disagree with a statement*. All other interviewees also indicated that their chosen endpoints covered their full spectrum of agreement/disagreement.

Insights into how some of the interviewees conceptually anchored their endpoints before choosing their verbal anchors, proved interesting. The prompts encouraged respondents to “Picture yourself talking to someone, and they say something that you agree with as much as you possibly could. Picture yourself responding to this person by saying “I ----- agree.” They were then prompted to think of a word to place before ‘agree’ forming the phrase, which would describe *the most they could possibly agree with a statement*. The same was done for the disagreeing verbal endpoint.

Several respondents talked about the scenarios they pictured when choosing their own personally-meaningful verbal anchor. Listed are some examples:

- Interviewee 6 discussed who she had imagined herself speaking to and said, “[Laughing] Talking to my mum! Probably because I’ve been calling her several times today.” She confirmed that she pictured herself saying “I absolutely agree” to something her mother might say. She felt she would equally use the word ‘absolutely’ in front of ‘disagree’ in a conversation with her mother.
- Interviewee 4 also said that he visualised a setting where he would typically express agreement; “I visualised my supervisor for my research project for some reason. [...] I would say [to him] “I definitely agree with that””.
- Interviewee 2 said he had pictured agreeing with a positive statement about his favourite team, Manchester United, and that he had pictured himself disagreeing with someone saying “all Muslims are terrorists”.

Where respondents chose to use different verbal anchors to represent their agreement/disagreement extremes (Interviewees 4, 15, 16, 2, 12, 14 in Table 4. 11), their reasons for doing so were explored in the retrospective debriefs. It was clear that

the verbal anchors they generated were personally meaningful to them, when you consider some of the reasons they gave;

Interviewee 16:

““I completely agree” doesn’t sound like something I would say, whereas, “I completely disagree” is something I would say.”

She indicated that both her chosen verbal endpoints represented the most she could possibly agree/disagree with something, however, she felt that she would naturally place different adverbs before ‘agree’ and ‘disagree’.

Interviewee 4:

“When I saw ‘agree’, - I just - ‘definitely’ sprung to mind. I think it’s the way I talk...maybe I associate ‘definitely’ with more positive things. And then, um, when I saw the ‘disagree’ side of things – I just – thought of another word really. I suppose ‘completely’ just sprung to mind, I don’t know if I associate that with being more firm and disagreeing...it was just my opposite.”

It is worth noting that all respondents who chose different verbal endpoints indicated that both of their endpoints (although different) represented, for them, their extremes on the agreement/disagreement cognitive continuum. This echoed the findings from Stage 1.

#### *4.4.5.8 Numerical conceptualisation*

In Stage 1, it was discovered that some interviewees desired numerically-imbalanced rating-scales, in that one side of the continuum had more intervals than the other. The IRSP software catered for this need, and some interviewees took up this option. See Interviewees 4, 16 and 14 in Table 4. 12.



**Table 4. 12 Numerical Endpoints Inputted by Respondents (Phase Two)**

IRSPv1						
Resp.	Verbal Endpoint (agree)	Verbal Endpoint (disagree)	Numerical Endpoint (agree)	Numerical Endpoint (disagree)	Modified Numerical Endpoint (agree)	Modified Numerical Endpoint (disagree)
1	totally	totally	8	-8		
4	definitely	completely	3	-3	3	-2
6	absolutely	Absolutely	2	-2		
8	strongly	Strongly	2	-2		
9	Fully	Fully	10	-10		
13	minimally*	slightly*	3	-3		
15	totally	highly	10	-10		
16	totally	completely	3	-4		

IRSPv2						
Resp.	Verbal Endpoint (agree)	Verbal Endpoint (disagree)	Numerical Endpoint (agree)	Numerical Endpoint (disagree)	Modified Numerical Endpoint (agree)	Modified Numerical Endpoint (disagree)
2	totally	fully	3	-3	2	-2
3	totally	totally	3	-3		
5	ok*	not ok*	6	-6		
7	totally	totally	5	-5		
10	Fully	fully	3	-3		
11	Agree*	Strongly	2	-2		
12	definitely	totally	4	-4		
14	definitely	completely	6	-5	5	-4

- a This respondent initially defined a numerically balanced IRS, yet subsequently modified it to a numerically-imbalanced IRS.
- b This respondent opted for a numerically-imbalanced scale.
- c This respondent initially defined a numerically imbalanced IRS, and subsequently modified it to make it shorter whilst still maintaining the numerical imbalance.
- \* The reason for these peculiar verbal anchors was explained in Table 1.

Interestingly, Interviewees 14 and 16 defined a numerically-imbalanced IRS right from the start, whereas Interviewee 4 initially defined his extreme disagree at ‘-3’ and his extreme agree at ‘3’, providing him with a numerically-balanced IRS with seven categories (a typical length with Likert formats). Interviewee 4, after completing Greenleaf’s items and examining his IRS using the graph visual aid, asserted that whilst he was using all of his varying levels of *agreement*, he did not think any *finer* than *two* stages when it came to rating his level of *disagreement* with something. Interviewee 14 also chose to modify her IRS (from -5-0-6 to -4-0-5), before proceeding to the main survey. The process of practicing the use of her IRS on Greenleaf’s items, led her to

conclude that whilst she did not need quite so many response categories, she still desired more “shades of grey” when ‘agreeing’ than when ‘disagreeing’.

This highlights another potential problem with respondents using researcher-defined fixed rating-scales to rate statements; it is clear that some respondents gradate their level of agreement to a greater/lesser extent than with disagreement. This has implications for the instrument validity of researcher-defined fixed rating-scales – given they are symmetrical – and lends further support to the argument for individualised rating-scales.

#### 4.4.5.9 *IRSP vs researcher-defined rating-scales*

Whenever interviewees compared the IRSP survey to surveys in general, this was entirely unprompted. However, there were some insightful comments which are mentioned next.

Interviewee 2 explained that he does not particularly enjoy surveys and does not like to think too much whilst filling them in. At one point he said “normally when you get a survey, it’s either one to five [referring to a typical fixed-rating scale]. And for me personally, I find that easier.” This raises the issue of whether respondents *want* the additional task of anchoring their own rating-scale, even if it is more *meaningful*. On this point, some interviewees had indeed chosen verbal or numerical anchors because they were familiar with them, having used them in general surveys:

- Interviewee 6 when asked why she chose  $\pm 2$ , she said she chose those numbers because she had felt comfortable using them in other surveys.
- Interviewee 8 said he chose the word ‘strongly’ “just because it’s used quite popularly in Likert scales all over the place, and it’s easy for people to

understand and it obviously emphasises that there's a firm belief that you agree with this or not".

However, the majority of the interviewees felt otherwise. Many *enjoyed* the personalised aspect of using their own IRS to rate survey items. For example, when asked "How did you find the defining-your-own-opinion-scale part?", Interviewee 4 responded,

"I thought it was quite good, yeah. It was nice to have a go yourself really. Because the surveys I have done in the past, just has a typical scale of its own...and you might not agree with the number of places they have on the scale, things like that. So it's quite good to have your own input."

Given the above insights, it was considered useful to see whether respondents would prefer to use an individualised rating-scale (IRS) or a fixed rating-scale. As such, it was deemed practical to include some measures in the quantitative phase that tested respondents' preference for using the IRSP method over typical survey rating-scales. These measures are discussed later in the quantitative chapter.

## **4.5 Stage 4: Pilot test**

### **4.5.1 IRSP Software Stress Test**

Before proceeding onto the quantitative phase of the study, a pilot test was necessary. However, before the pilot test was conducted, the programmer was asked to perform a stress test to check whether the survey software and the database structure could handle multiple respondents completing surveys simultaneously. It was important to ensure that data would not be lost as a result of high respondent traffic. The stress test was created by simulating about 500 data connections to the web service that is responsible for retrieving survey information and saving survey responses. The connections were run

from a single computer that simulated multiple, non-sequential connections to the web service by using threads. This helped to simulate a more realistic scenario of multiple respondents accessing the web service at the same time. No response data was lost in the process. Thus, the survey system survived the stress test.

#### **4.5.2 Method**

The main objective of the pilot test was to provide a live test of the IRSP survey in what was likely to be a similar response environment to that of the main quantitative test. The main issues that the pilot test sought to address were the following:

- To make a final comparison of both versions of the IRSP, and confirm whether the IRSPv2 (the likely choice) still performs better, when compared to IRSPv1;
- To make additional checks on the performance of the survey software;
- To use a live test to confirm the results from the survey stress test; that the platform can handle multiple respondents completing the survey simultaneously and still capture the data accurately.
- To report any issues flagged with particular items.

##### *4.5.2.1 Online survey*

It was important for the pilot test to simulate the data collection process that would be used in the main phase of quantitative research. As such, respondents were sent an email containing the survey link. They commenced the survey by clicking on the link and followed the instructions presented to them on the web page. The online survey was similar to that used in the CVP-RD interviews (after the aforementioned amendments). Naturally, all instructions relating to the CVP-RD process (such as the ‘think aloud’ instructions) were not present, but other instructions appropriate to *this* test were added.

To see the screenshots for both versions of the survey (i.e. IRSPv1 and IRSPv2) used in the pilot test, see Appendices J and K.

The variables measured were the same as those in the CVP-RD interviews: the demographic items, the Cognitive Style Indicator (CoSI) scale, the Affective Orientation (AO) scale, the Personal Need for Structure (PNS) scale, the Big Five Inventory (BFI) scale and the Mood indicators. No changes were made to item-wording or the order of items.

#### 4.5.2.2 *Sample*

The pilot test was conducted with Bradford University full-time MBA students. This sample was used for three key reasons:

1. The MBA students are a demographically and culturally diverse group of students, which would ensure that reactions to the survey content would come from a wide range of perspectives.
2. The students enrolled on this module were required to take part in weekly lab classes held on-site in designated computer rooms, and supervised by a lab tutor. This made for an ideal setting, as a large number of students could be observed completing the survey simultaneously and feedback would be instant.
3. Their email addresses were known and they could be targeted separately to the rest of the student population. This meant that participation in the pilot-test would be a controlled separate experiment on a small sub-group of the larger student population.

#### 4.5.2.3 *Setting and procedure*

The MBA students were approached at the start of a lecture, one week prior to the pilot test. They were told that at the start of their next lab class they would be given the option to take part in voluntary online survey. It was explained that it would not be a compulsory element of their lab class, but if they chose to participate it would be useful for two reasons. Firstly, they would be able to see first-hand how other post-graduates engage in primary data collection, which could be helpful to them on designing their own research projects. Secondly, personalised feedback would be emailed to every participant, alongside a summary of scores and an explanation of how they are interpreted. Additionally, they were informed that should they wish to spare only five minutes of their time, there would be an exit point five minutes into the survey where they would be invited to exit should they so desire (but that individualised feedback could not be provided to those who exited early). Those, however, who chose to proceed onwards with the rest of the survey, would benefit from the feedback promised, and that it would take approximately fifteen minutes. The exit point page appeared *after* participants defined their IRS and after using it to rate Greenleaf's items, but *before* the psychometric scales. This was done so that, at the very least, volunteers who may not have wanted to spare the full fifteen minutes would still participate in the earlier section of the survey, which is the key part.

On the morning of the day of the lab classes (for which there were six), an email was sent to the MBA students containing some instructions and two html links. To see a copy of this email refer to Appendix L. The six lab classes were assigned to particular versions of the IRSP survey, shown in Table 4. 13.

**Table 4. 13 Pilot Study with MBA Students: IRSP versions and lab classes.**

	<b>Lab Class 1</b>	<b>Lab Class 2</b>
<b>11am-1pm</b>	IRSPv1	IRSPv1
<b>2pm-4pm</b>	IRSPv2	IRSPv2
<b>4pm-6pm</b>	IRSPv1	IRSPv2

This enabled both versions of the survey to be tested with a maximum amount of respondents completing it simultaneously (i.e. a live stress test of IRSPv1 during the 11am-1pm class, and of IRSPv2 during the 2pm-4pm class). The lab classes from 4-6pm provided a test for whether two different surveys (IRSPv1 and IRSPv2) could also be simultaneously filled in by multiple respondents, without problems in data capture.

Obviously, the researcher could not be present in all the lab classes. Therefore, to ensure that respondents in all classes received the same set of instructions before starting the survey, a lab-tutor instruction-sheet was given to the second lab tutor. Both the researcher and the other tutor read these instructions to the lab class at the start of the session (see Appendix M for the lab-tutor instructions), which ensured a consistent method of instruction-delivery. The lab-tutor instructions, in short: thanked everyone who chose to participate; asked that everyone make sure they each have a sheet entitled “Instruction Sheet” and that they read through it before clicking onto the survey; stressed that people open and read the participation email carefully, and that they click on the correct survey link *appropriate* to them (given their lab slot); and finally that they complete the details on the “Instruction Sheet” and hand it back at the end of the survey. The “Instruction Sheet” given to participants (see Appendix N) reminded them that there would be an exit point after approximately five minutes, but that personalised feedback could only be provided to those who completed the entire survey.

Additionally, it requested four details which the respondents had to complete and hand in to the lab tutor after the exercise. The details requested were the following:

- They had to tick one of two boxes, indicating whether they had completed the ‘short version’ (i.e. exiting early) or the ‘entire survey’.
  - This was useful because their manual report would need to corroborate the data in the database. For example, should someone have ticked ‘entire survey’, yet their data record in the database was incomplete, this would have highlighted a potential problem with the software’s capture of the data.
- A large box was provided where they were able to note down any words that they found ambiguous or confusing, as well as any other comments.
  - This would help to further scan the scales for likely problem-items.
- They were asked to provide their student number and surname.
  - This was done so that their feedback could be paired up with their survey entry in the database.

Approximately four weeks after taking part in the pilot tests, the participants were emailed individual feedback on their BFI and CoSI scores with an interpretation, provided by the academics that developed the scales, and a document summarising the problems that had been highlighted in the survey pilot, with a list of useful amendments made as a direct result of their involvement (see Appendix O).



### 4.5.3 Findings

#### 4.5.3.1 Sample

In total, there were twenty-seven respondents who completed the IRSPv2 survey, all of whom completed it in its entirety (i.e. no one opted out at the first exit point) and twenty-four respondents completed the IRSPv1 survey, with only one person completing only the first part (i.e. they opted out at the first exit point). Of the twenty-seven that completed the IRSPv2 survey, nineteen were male and eight were female. Of the twenty-four who completed the IRSPv1 survey, eighteen were male and six were female. Therefore in both groups, approximately three quarters of the sample were male, consistent with the cohort demographic. Their age spread in the sample is shown in Figure 4. 34. Most respondents were in their late twenties.

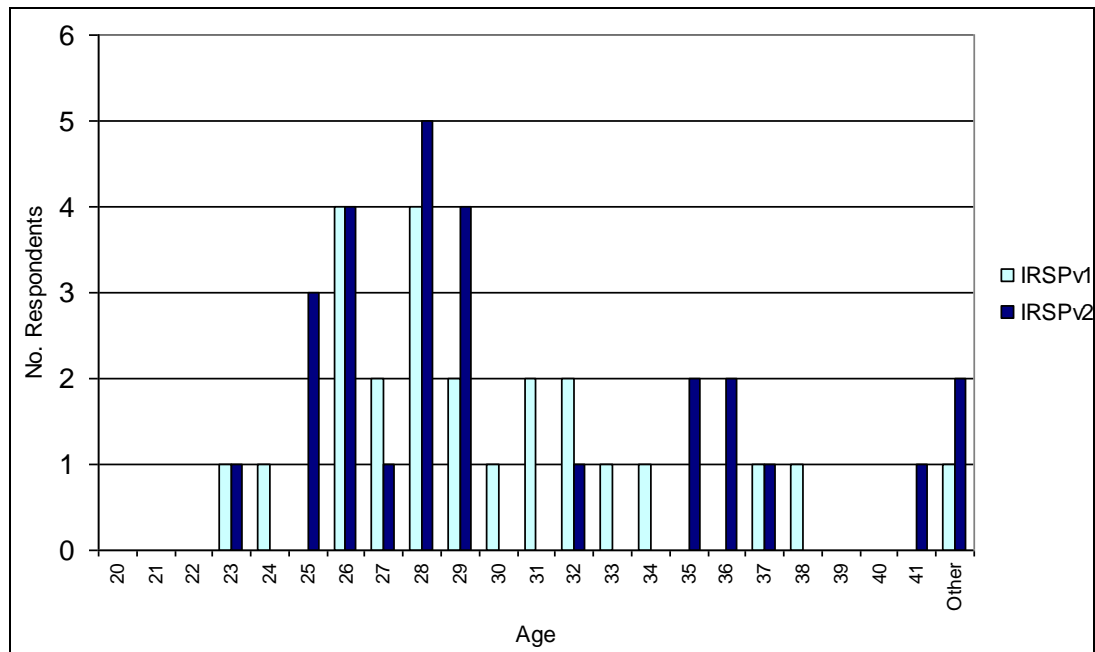


Figure 4. 34 MBA Pilot: Age spread

Respondents in both IRSPv1 and IRSPv2 groups represented a wide range of cultural backgrounds, as evidenced by their ethnic spread (Figure 4. 31 and Figure 4. 32) and their first languages (Figure 4. 33 and Figure 4. 34).

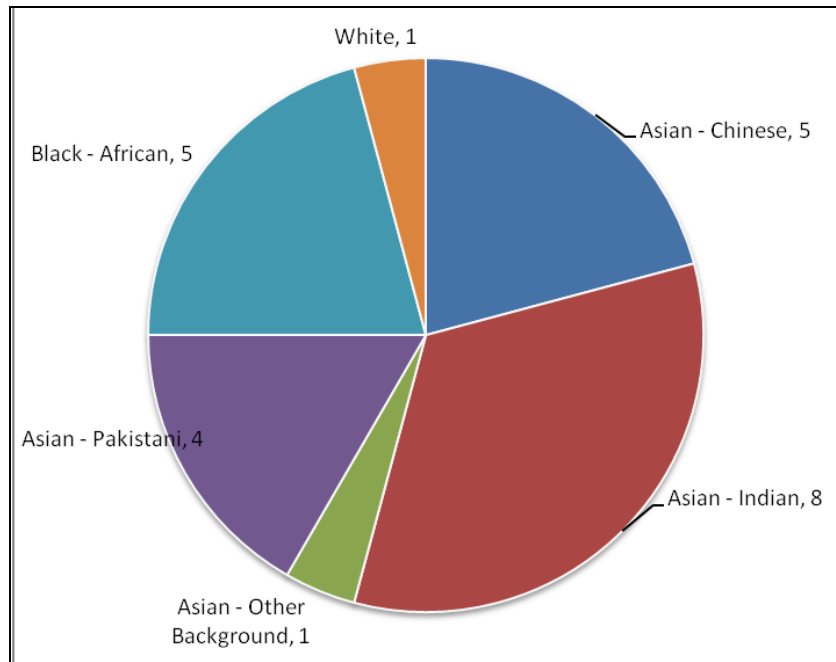


Figure 4. 35 MBA Pilot IRSPv1: Respondents' ethnicities.

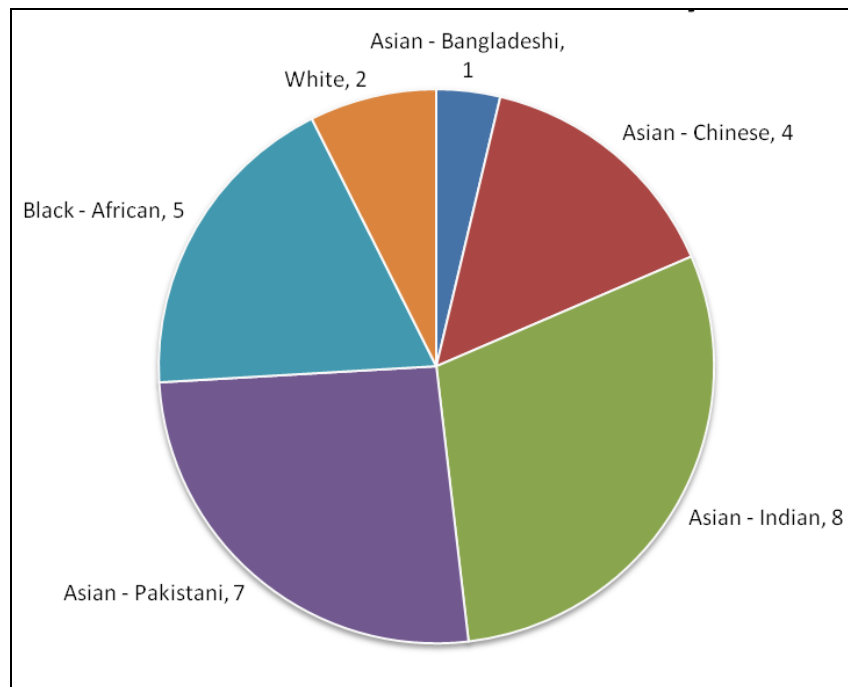
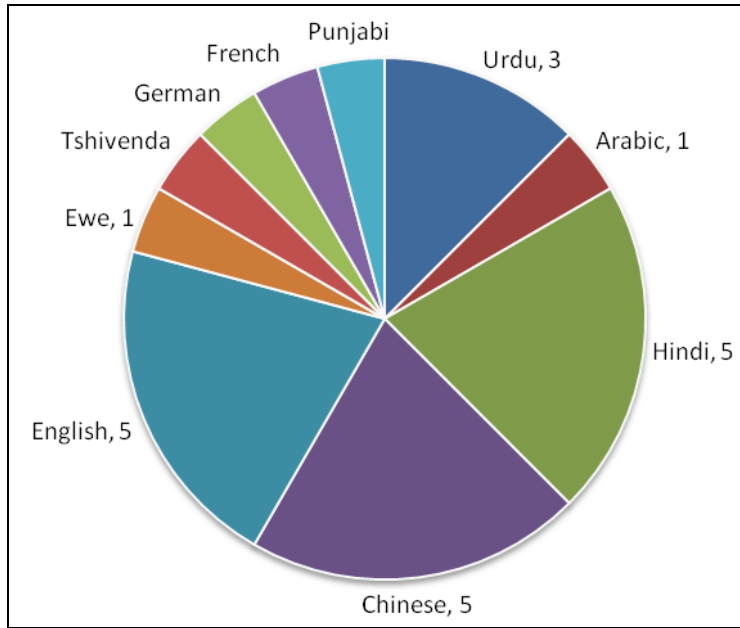
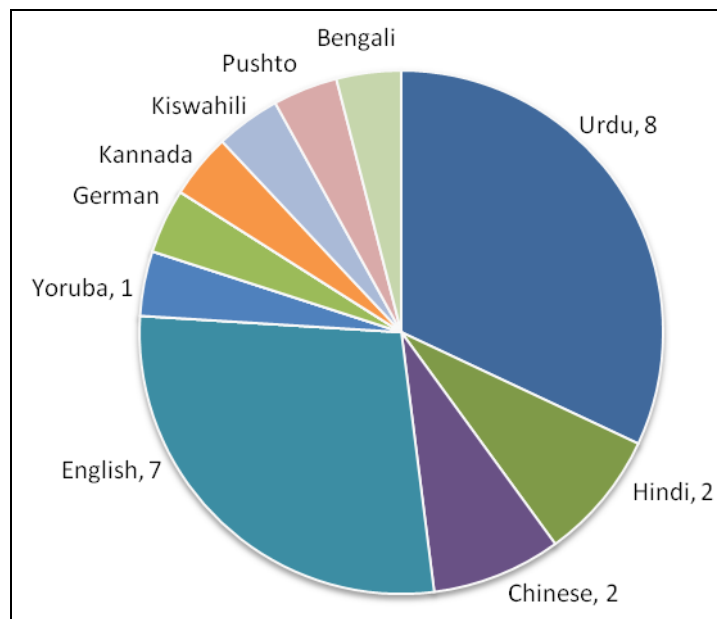


Figure 4. 36 MBA Pilot IRSPv2: Respondents' ethnicities.



**Figure 4. 37 MBA Pilot IRSPv1: Respondents' first language.**

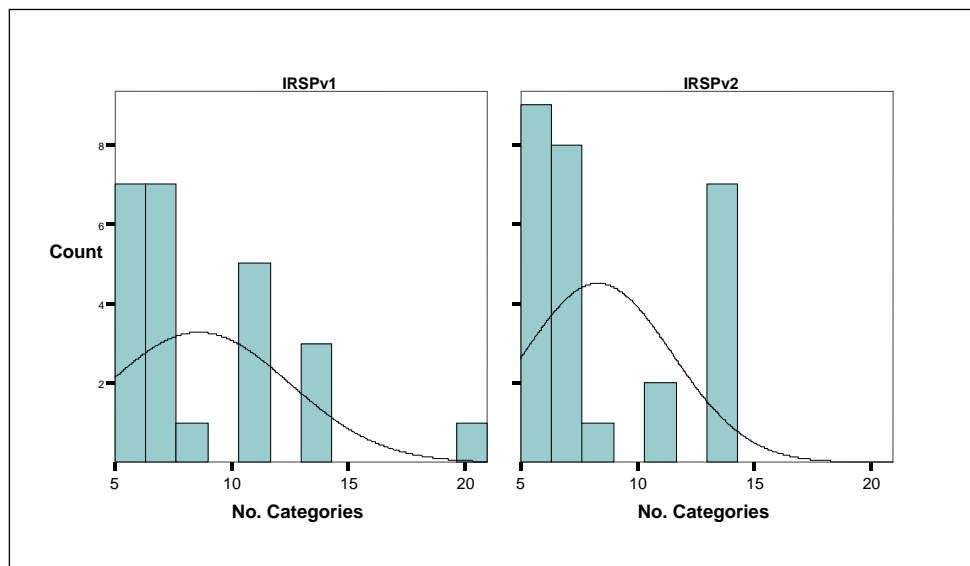


**Figure 4. 38 MBA Pilot IRSPv2: Respondents' first language.**

#### 4.5.3.2 IRSPv1 vs IRSPv2

When examining the lengths of the IRSs chosen by respondents (i.e. the number of categories on their rating-scales) across the two groups, the mean number for those who completed IRSPv1 was 8.63 (s.d.=3.899) and 8.33 (s.d.=3.187) for those who had

completed IRSPv2. Whilst the mean and standard deviation appear marginally shorter for IRSPv2, the difference is not significant. Figure 4. 39 shows the spread of number of categories chosen by respondents for both groups. Similar to Stage 3, IRSPv2 appears to dissuade respondents from the mystery attraction to  $\pm 10$ , and respondents appear to choose shorter rating-scales. However, this difference is not as markedly significant here as it was in Stage 3.



**Figure 4. 39 MBA Pilot: Histograms showing no. categories chosen by IRSP group.**

It was useful to calculate the index of dispersion for all the respondents, as it provides a measure of the evenness of use of response categories (Wyer, 1969). Krishnamurty et al. (1995: 290) define index of dispersion as “the proportion of dispersion that exists within the observations relative to the maximum dispersion that can possibly exist.” Index of dispersion (I) scores ranges from 0 to 1. A score of ‘0’ would occur if a respondent only used one of their response categories. A score of ‘1’ would occur if all intervals on a rating-scale were used equally (i.e. a uniform distribution). As per Krishnamurty et al. (1995), the following calculation was used to compute index of dispersion:

$$D = \frac{k(n^2 - \sum fi^2)}{n^2(k - 1)}$$

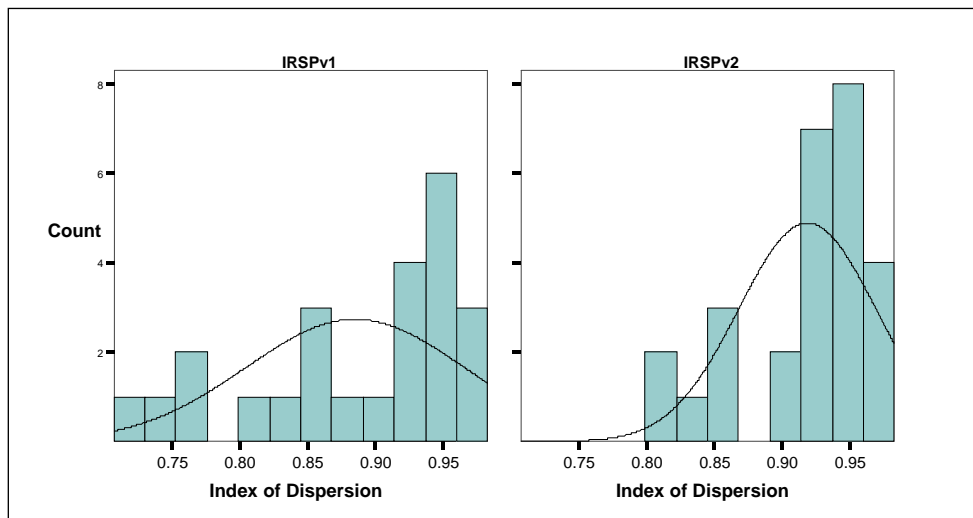
Where,

*k* = number of categories

*n* = total number of observations

*fi* = frequency in each category

When comparing the mean index of dispersion and the standard deviation for those who completed IRSPv1 (I=.885, s.d.=.081) and those who completed IRSPv2 (I=.918, s.d.=.051), those who completed the IRSPv2 survey appear to have more even responses. However, a t-test did not find the difference statistically significant. When examining histograms showing the spread of the index of dispersion scores in both groups (see Figure 4. 40), IRSPv2 appeared to perform better, given the spread is skewed more to the right, and thus skewed more closely to 1.



**Figure 4. 40 MBA Pilot: Histograms showing index of dispersion by IRSP group.**

On the whole, those completing the IRSPv2 survey appeared to spend less time pondering the numerical anchoring of their IRS. In a post-pilot discussion with some of

the MBA students, one of the participants had mentioned that he had wanted to have an IRS of  $\pm 3$  but found the IRSPv1 instructions somewhat unclear and had accidentally assigned himself an IRS of  $\pm 2$ . The quantitative evidence was not compelling enough on its own to suggest that the IRSPv2 be chosen. However, there is no evidence to suggest IRSPv2 is *inferior*, and the qualitative insights both from Stage 3 and 4 suggested that choosing to carry forward the IRSPv2 presented clear advantages.

#### 4.5.3.3 *Survey software performance*

Overall, the IRSP survey software performed well. However, the pilot permitted the discovery of a variety of small issues that were resolved before the next stage:

- Many respondents experienced difficulty because there was no option to go back to previous pages in the survey. Some clicked the ‘back’ button on the browser, only to find that they were taken back to the start of the survey. This was frustrating for those who had encountered this.
  - Solution: It would have been very costly and problematic to have the programmer create a software amendment to permit back-tracking in a survey. The benefit of this did not outweigh the cost (as it was obvious that very few respondents were likely to feel the need to revisit previous pages). Thus, adding a ‘back’ button to the survey pages was not a viable option. However, a modification was made to the software so that when a respondent clicks on a survey link, it opens in a new browser window, with the ‘back’ button disabled. This would prevent respondents from clicking the ‘back’ button on the browser. The only flaw with this occurs where respondents copy and paste the survey link into an existing

browser window (with existing browser history), as the ‘back’ button cannot be disabled in this scenario. In order to account for the few that do it this way, a message was featured in the next version of the survey, asking that respondents do not use the ‘back’ button.

- One demographic question was found to be ambiguous, the ‘total number of years at university’ question.
  - Solution: The wording for the ‘total number of years at university’ was changed so that it was clearer; “Total number of years you have spent so far in university education (includes current and any previous courses)”.
- Several participants noted that the demographic question on national identity was partially hidden (when placing a tick in ‘other’).
  - Solved: The scroll bar was lengthened so that this question could be seen properly.
- Some participants expressed that they would have liked to be able to modify their verbal endpoints on their IRS, given that the survey only permitted them to modify their numerical endpoints.
  - Solution: This feature was introduced. The original inputs were ghosted into the boxes to help respondents remember what they had selected first time around.
- Some felt that if the graph had a few extra labels it would be clearer.
  - Solution: Titles were added to the x- and y-axis, and a hover-label was added for when the mouse cursor passes over the bars, with a message saying, for example, “3 was used 4 times”.

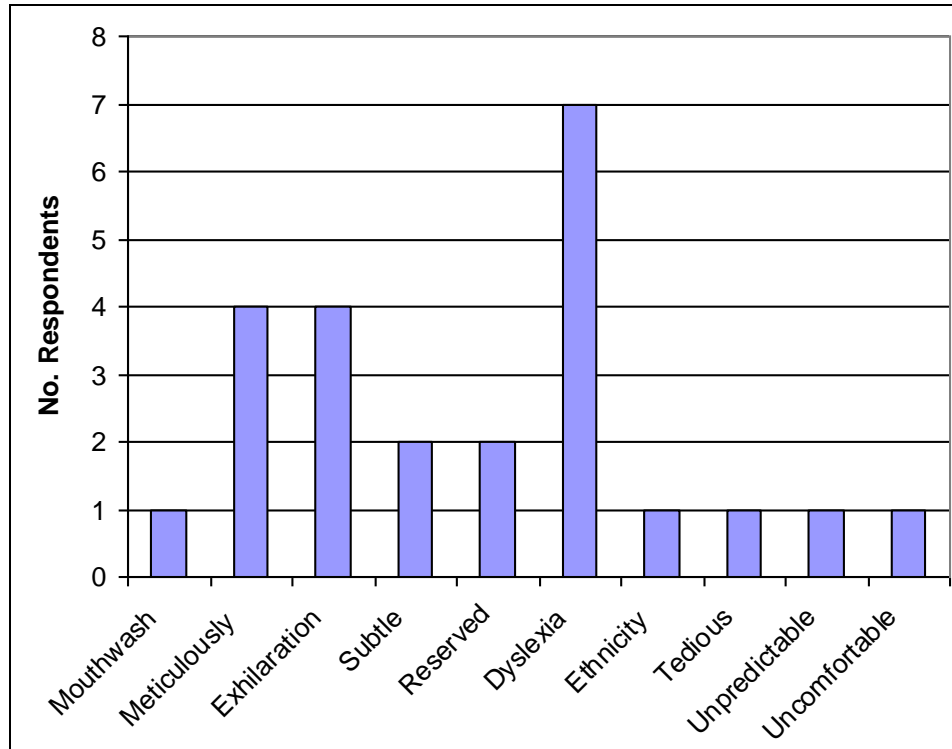
- A few participants indicated that they would have preferred not to have had to give their ‘name’ as part of the demographic information, with some indicating that it might make them less likely to be honest about their opinions.
  - Solution: ‘Name’ was removed from the demographics page to encourage respondents’ trust in the anonymity of responses, and thus encourage them to provide more honest answers.

Data capture was perfect, and the surveys passed the live stress test easily. However, there was a big error spotted in the *presentation* of the data captured. The programmer designed a ‘researcher interface’ database that strung together the data in a way that made it user-friendly for the researcher. However, there was only one record per participant in this database, which was a peculiarity, given there were a small handful of students that had filled in a survey on more than one occasion, as they clicked ‘back’ and had had to start the survey again. The problem was easily fixed, by switching off a filter that disallowed multiple records with the same demographic data from appearing in the ‘researcher interface’ database. These missing records reappeared and the issue was solved.

#### 4.5.3.4 *Problem items*

All the “Instruction Sheets” were returned completed. The comments reported were summarised (see Appendix P). Any problem-words, that they had listed, were tallied and their corresponding item name was also noted. Figure 4. 41 shows the words that were flagged, and the number of respondents who flagged them.





**Figure 4. 41 Problem-words flagged by MBA students in their feedback sheets.**

It is worth stressing that these problem-words were considered in the context that over three quarters of respondents in the sample did not speak English as a first language. ‘Dyslexia’ was the most commonly misunderstood word, with ‘meticulously’ and ‘exhilaration’ coming in joint second. It is likely that the majority of people in the U.K. know that dyslexia is a learning disability, given it is frequently referred to throughout school and higher education. However, many of the MBA students were not familiar with it. The decision was made to keep this demographic item in the survey because those that do not know what it is, are unlikely to select ‘yes’ when asked whether they have been diagnosed with dyslexia.

Figure 4. 42 shows what items the problem-words belong to. Notice that four of the bars have been coloured with horizontal stripes. These bars are of particular interest because

those very same items (and very same words) had been flagged in Stage 3 during the CVP-RD interviews.

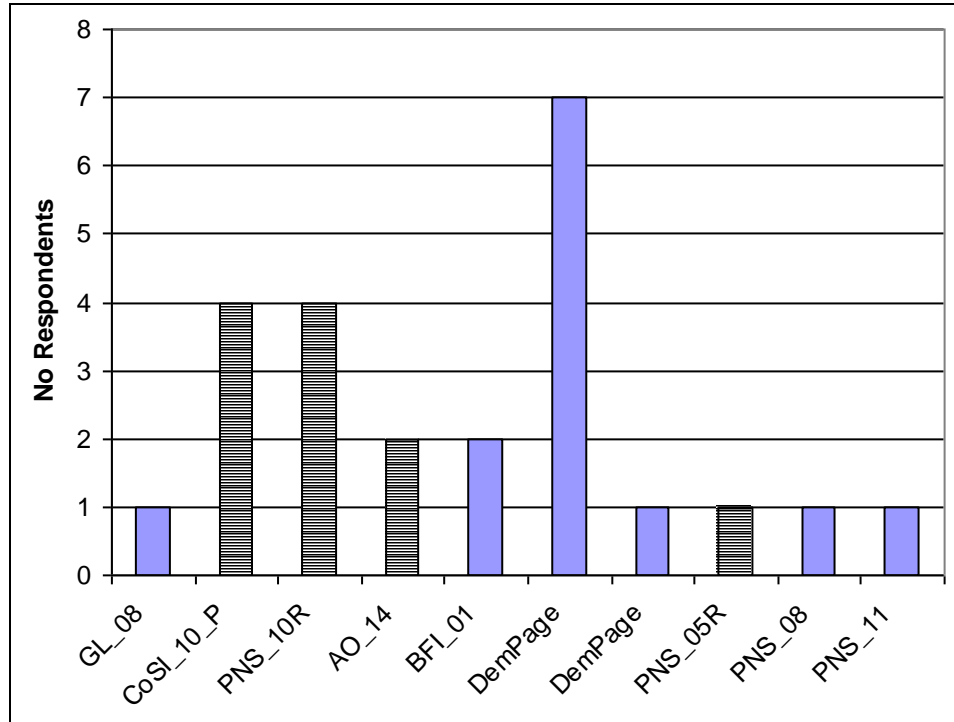


Figure 4. 42 Items to which the problem-words belong.

Whilst this was worthy of note, the items in question were *not* removed from the survey. This decision was taken because, had items been removed, the measurement models for each of the psychometric constructs would have been invalidated, given their validity had been pre-established in their current form. It was still useful, however, to have this background understanding about some potential problem-words as it could later explain peculiar item loadings.

#### 4.6 Summary

This chapter presented the four key stages that led to the development of the IRSP Survey Software, which was subsequently tested in an online large scale quantitative

study. Stages 1 and 3 outlined how the qualitative interviews collected led directly to development, and provided the underlying insights about respondent cognitions. Stage 2 detailed the steps taken to transform the IRSP from ‘paper’ to ‘computer’, through software design. Stage 4 described the quantitative pilot test conducted, and final adjustments.

Stage 1 started with an outline of the foundations upon which the original design for the rudimentary IRSP was based. It detailed the initial decisions behind the creation of the visual aid, instruction order and wording. The findings of each of the four rounds of data collection from Stage 1 interviews were presented, along with other observations. The inability of interviewees using the IRSPr1 instruction sheet (Round 1 interviews) to define meaningful intervals (and a predisposed attraction to numerical anchors of ‘10’, and multiples thereof) prompted further development to the IRSP: the instruction wording was simplified; an instruction was added to clarify that respondents could choose same or different verbal labels; and Greenleaf’s sixteen uncorrelated items were included in the exercise.

After Round 2 in Stage 1, respondents were still attracted to ‘10’ as a numerical endpoint, which meant meaningfulness of intervals was not always achieved. Before proceeding to Round 3 interviews, the instructions were modified further: wording in some of Greenleaf’s items were adapted for use with British students; the definition of the neutral position was modified; an instruction and title were added, informing respondents of the purpose of the exercise; the visual aid was separated from the instructions and improved; the term ‘adverb’ was not included in the instructions, given a number of students were unclear of its definition; numerical-anchoring instructions

were modified and an example was included. Round 3 of the interviews indicated that the numerical-anchoring *example* had had a biasing effect. As such, it was removed from the instructions and further modifications were made. Interviews from Round 4 indicated that the IRSP<sub>r4</sub> instructions had succeeded in better dissuading respondents from choosing  $\pm 10$  for their endpoints, and had them choose more meaningful rating-scales. Furthermore, insights showed that the interviewees enjoyed choosing their own verbal labels, and would not have wanted to be provided with a list of adverbs as an aid. The verbal anchoring instructions were modified in line with insights gained.

Round 4 of Stage 1 also heralded the end of the IRSP development on paper, with all planned modifications saved for the electronic version. The facility to provide respondents with an option to modify their IRS was planned for, to circumvent the issue of respondents defining IRSs that did not represent their ideal rating-scales. The planned facility included a bar chart to aid respondents and the option to remove it entirely if so desired. A second version of the IRSP was developed (IRSP<sub>v2</sub>), based on insights from the interviews. Interviewees reported they were able to define IRSs with ease, and that the task was not difficult. Insights suggested that respondents conceptualised *agreement* and *disagreement* as either bipolar or unipolar, and their choice of verbal and numerical anchors reflected this.

Stage 2 outlined the steps taken to inform the creation of the IRSP software. It also summarised its key features.

Stage 3 detailed an additional phase of qualitative data collection, through sixteen concurrent verbal protocol-retrospective debrief (CVP-RD) interviews. The IRSP

Survey Software performed well, although minor modifications were made. The IRSPv2 performed better than the IRSPv1 in having respondents define more meaningful IRSs. The graph page was found to be a very useful part of the IRSP, in aiding respondents to reflect on the meaningfulness of their IRS, and modify it if necessary. Furthermore, the graph page appeared to render respondents conscious of any response style behaviour and encourage them to reflect their opinions more accurately. Some interviewees experienced difficulty with some of the wording of certain items. These items were noted so that they could be inspected in the main quantitative phase of the study. Insights suggested that the verbal labelling process successfully resulted in respondents capturing their entire agreement/disagreement continuum, with personally meaningful labels being chosen. For some respondents there was a clear difference between the number of subjective categories they had for 'agreement' and 'disagreement', providing a further justification for the practical use of the IRSP in improving measurement validity. A question of whether respondents would prefer to continue using researcher-defined (fixed) rating-scales was raised. Even if IRSs proved to be more meaningful for respondents, and ultimately resulted in an improvement in data quality, if the IRSP was considered to be burdensome by respondents then its usefulness would be redundant. As such, provisions were made to investigate this in the main quantitative phase of testing.

Stage 4 involved a pilot test with MBA students. The IRSP Survey Software passed a stress test, indicating it was robust to multiple users. The pilot test was conducted in the manner planned for the main quantitative test, so that it provided a realistic pilot. As such, respondents completed the survey online, having had a link sent to them by email. IRSPv2 emerged as the best version of the IRSP to carry forwards into the next

quantitative phase of testing. A few final modifications were made to the IRSP Survey Software, based on insights from the pilot test.

## **Chapter 5. Testing the Individualised Rating-Scale Procedure**

---

## 5. Testing the Individualised Rating-Scale Procedure (IRSP)

### 5.1 Introduction

This chapter addresses the second part of the research objective:

*To **test** a measurement instrument capable of having respondents individualise their own rating-scales for use in online surveys.*

This chapter is divided into four stages. The first three stages directly address the above research objective. The fourth stage explores what might predict rating-scale choices. Stages 1 and 2 see whether the IRSP was able to capture respondent data in an equivalent manner to the Likert-type rating-scale (LTRS). Questions about whether person-specific characteristics are linked to the types of IRSs respondents define are redundant, if at the very basic level the IRSP is not operating adequately as a measurement instrument. As such, the quantitative analysis started with an exploration of whether data captured using the IRSP fit the measurement models equivalently well to the data captured using the pre-validated (for each construct) Likert-type rating-scales (LTRS). To achieve this aim, structural equation modelling (SEM)<sup>1</sup> was employed, through software packages AMOSv7 and SPSSv15. It was particularly suited given that the measurement models being tested were already pre-established, and the aim was to *confirm* whether the observed data captured, by both IRSP and LTRS, fit the measurement models adequately. As such, confirmatory factor analysis (CFA) was used. Stage 3 consisted of a formal test of reliability and validity of the IRSP, by comparing its test-retest reliability and validity to that of the LTRS. To this aim, the degree to which the data (split by measurement method used) in time period 1 (T1) was

---

<sup>1</sup> SEM is a family of statistical models that seek to explain the relationships among multiple variables, expressed in a series of equations, similar to a series of multiple regression equations. SEM combines interdependence and dependence techniques, and its foundation lies in two familiar multivariate techniques: factor analysis and multiple regression analysis (Hair et al, 2006). SEM is known by many names: covariance structure analysis, latent variable analysis and confirmatory factor analysis (Hair et al., 1998).



replicated in time period 2 (T2), was tested. In stage 4 individual characteristics are investigated for their relationship with the type of IRS chosen, in terms of length and numerical balance. In addition, descriptive statistics illustrating verbal labelling preferences are presented. Finally, respondents’ feedback on the IRS in comparison with the LTRS, are outlined.

**5.2 Stage 1: Establishing Model Fit**

*Objective: To assess measurement model fit for both the IRSP and LTRS groups; loose cross-validation.*

The online survey achieved a large number of responses, with 1,363 complete responses in T1, and 994 of those returning to complete the re-test in T2. Table 5. 1 shows the figures obtained broken down by test group, along with the experimental mortality figures.

**Table 5. 1 Sample sizes for groups in T1 and T2**

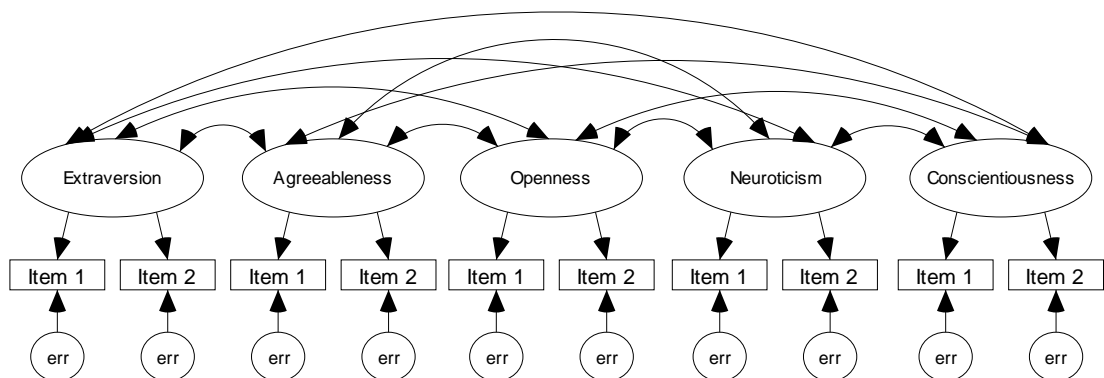
Test Group	Treatments	No. respondents that completed T1	No. respondents that completed T1 & T2*	Experimental Mortality	Test-retest %
TG <sub>1</sub>	IRSP-IRSP	386	282	104	73.1
TG <sub>2</sub>	IRSP-LTRS	393	297	96	75.6
TG <sub>3</sub>	LTRS-IRSP	293	202	91	68.9
TG <sub>4</sub>	LTRS-LTRS	291	213	78	73.2
<b>Totals</b>		<b>1363</b>	<b>994</b>	<b>369</b>	<b>Avg. = 72.9</b>

\*Information concerning T2 is included for completeness, even though T2 data was not required to achieve the objective of Stage 1 of the analyses.

The four psychological constructs measured by the online survey (Need for Precision – PNS, Big Five Index-10 – BFI, Cognitive Style Indicator – CoSI, and Affective Orientation – AO) have been previously validated. Therefore, all have predetermined measurement models consisting of the construct of interest (an exogenous variable) and their respective reflective indicators (each of the scale items). However, even with a

well established scale, it is important for the researcher to confirm its validity and dimensionality (Hair et al., 2006).

There are various modelling strategies employed through the application of SEM. Considering that the objective of this stage was to *confirm* whether the previously validated measurement models *fit* the data, it was necessary to adopt a *confirmatory modelling strategy* (as per Hair et al., 2006). As such, the single measurement model for each of the constructs was specified in AMOS 7, and SEM was used to see how well each model fitted the data. The BFI-10 construct was not examined at this stage given that it consists of multiple under-identified (fewer than three items) variables, exhibited in Figure 5. 1.



**Figure 5. 1 Big Five Index 10-item**

Before proceeding further, the data was checked for outliers.

### 5.2.1 Outliers

Outliers can be problematic when they are not representative of the population as they can seriously distort statistical tests (Hair et al., 1998). Specifically, outliers can have dramatic effects on the indices of model fit, parameter estimates, and standard errors

(West et al., 1995). As such, it was very important to examine the data for potentially harmful outliers.

Two methods were used to identify outliers: (a) the SPSS anomaly detection procedure that identifies unusual cases based on deviations from the norms of their cluster groups; (b) the multivariate detection of outliers in AMOS, which utilises the Mahalanobis  $D^2$  measure of the distance in multi-dimensional space of each observation from the mean centroid of the observations. Both methods highlighted the same extreme cases as potential outliers.

The SPSS procedure produced: peer groups based on a clustering model that explains natural groupings within the data; peer group norms for the continuous variables measuring AO, CoSI and PNS; anomaly indices based on deviations from peer group norms; and variable impact values for variables that most contribute to a case being considered unusual. The scatterplots (Figure 5. 2, Figure 5. 3, and Figure 5. 4) chart the respondents' variable impact measure (their most unusual item score) against its corresponding anomaly score. Annotated, are the cases that on closer inspection, presented too much deviation from the group to be included in further analyses, and were therefore removed before proceeding. Following each scatterplot is a small table presenting the scores of each of these outliers (Table 5. 2, Table 5. 3, and Table 5. 4). The tables also include the Mahalanobis  $D^2$  scores and the corresponding p-values, and are ordered from highest to lowest.

When examining the AO data (Figure 5. 2, Table 5. 2), case 52 was not flagged as an outlier by the SPSS test, but it had a particularly significant Mahalanobis score from the

multivariate test in AMOS (Table 5. 2), indicating it was a multivariate outlier. Given the results, cases 24, 52, 75, 247, 347, and 1158 were removed from further analyses on the AO data.

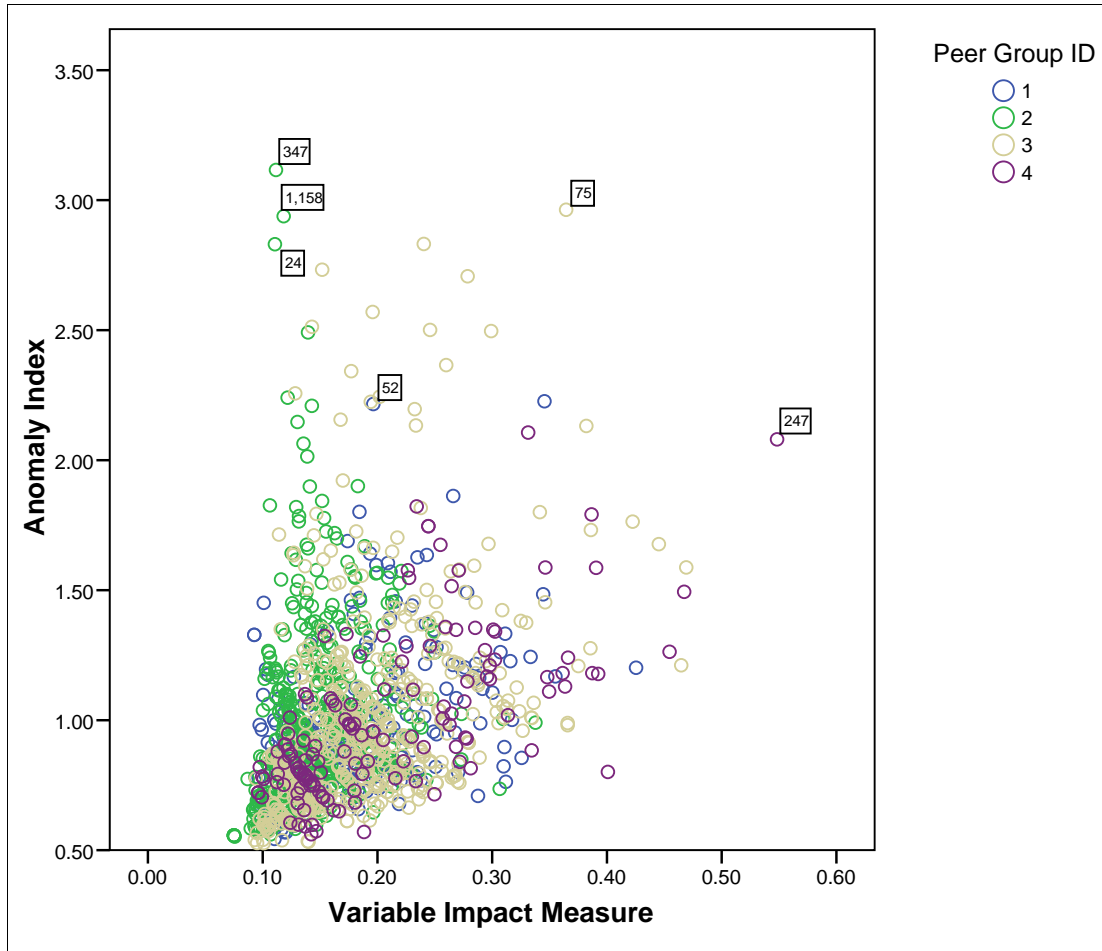


Figure 5. 2 Scatterplot AO: Illustrating outliers through respondents’ anomaly indices.

Table 5. 2 AO: Unusual cases’ anomaly indices and Mahalanobis scores.

Respondent	Anomaly Index	Variable Impact	Mahalanobis D <sup>2</sup>	p1*	p2*
1158	2.938	.118	111.043	.000	.000
52	2.22	.200	83.594	.000	.000
347	3.116	.112	78.127	.000	.000
75	2.963	.364	65.334	.000	.000
24	2.830	.111	61.816	.000	.000
247	2.080	.550	39.961	.000	.000

\* The p1 column shows, assuming normality, the probability of D<sup>2</sup><sub>i</sub> exceeding the given D<sup>2</sup>. The p2 column shows, still assuming normality, the probability that the largest D<sup>2</sup><sub>i</sub> would exceed the given D<sup>2</sup>. Small numbers in the p1 column are to be expected. Small numbers in the p2 column, however, indicate observations that are improbably far from the centroid under the hypothesis of normality.

On examination of the PNS data (Figure 5. 3, Table 5. 3), cases 230, 568, and 858 were removed from the PNS data.

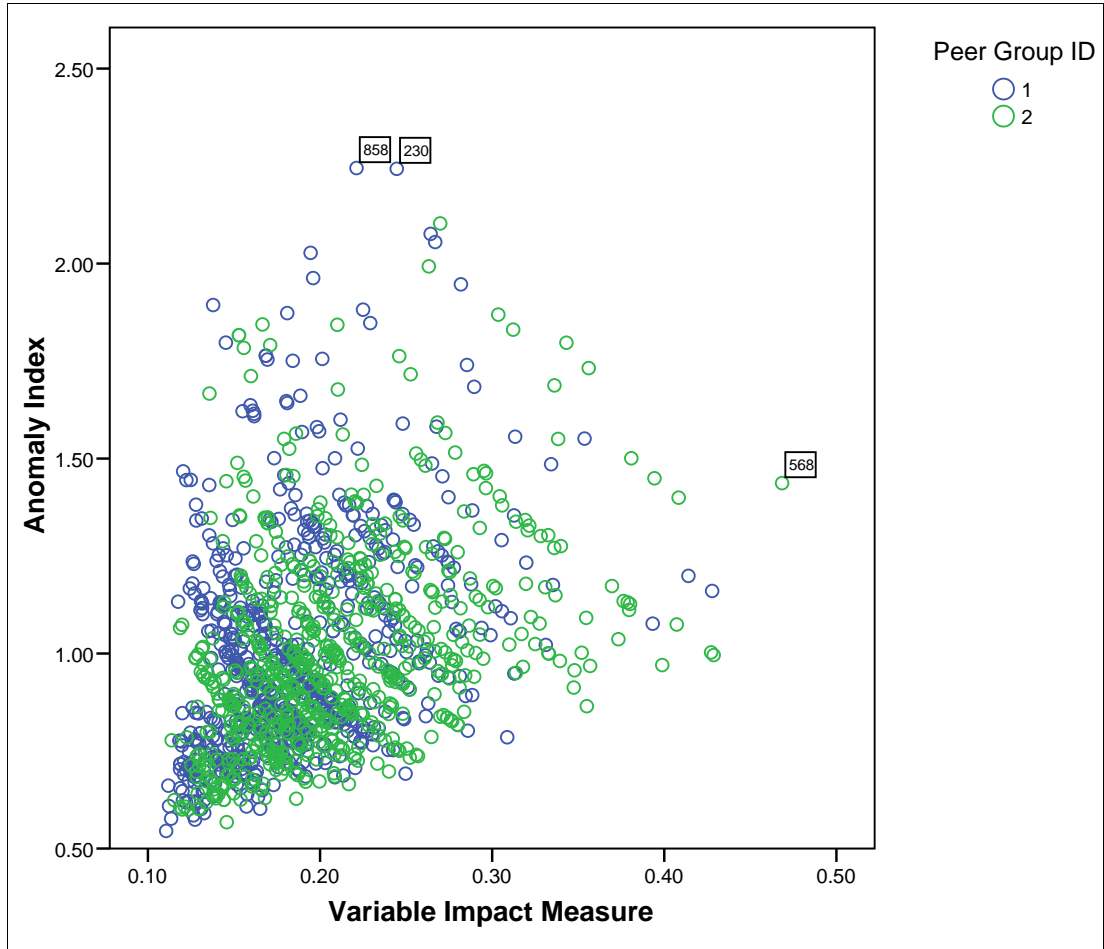


Figure 5. 3 Scatterplot PNS: Illustrating outliers through respondents’ anomaly indices.

Table 5. 3 PNS: Unusual cases’ anomaly indices and Mahalanobis scores.

Respondent	Anomaly Index	Variable Impact	Mahalanobis D <sup>2</sup>	p1	p2
230	2.243	.245	54.195	.000	.000
858	2.245	.221	34.810	.000	.000
568	1.44	.470	25.563	.008	.000

On examination of the CoSI data (Figure 5. 4, Table 5. 4), respondent 61 was not identified as an outlier by the Mahalanobis test in AMOS, but the scatterplot illustrated that it differed significantly from the rest of its peer group using SPSS’ univariate test, and it possessed a high variable impact score. Whilst cases 1125 and 301 were not flagged as outliers by the SPSS test, they had particularly significant Mahalanobis

scores from the multivariate test in AMOS (Table 5. 4), indicating they were multivariate outliers. Given the results, cases 61, 231, 301, 485, 1125, and 1130 were removed from the CoSI data.

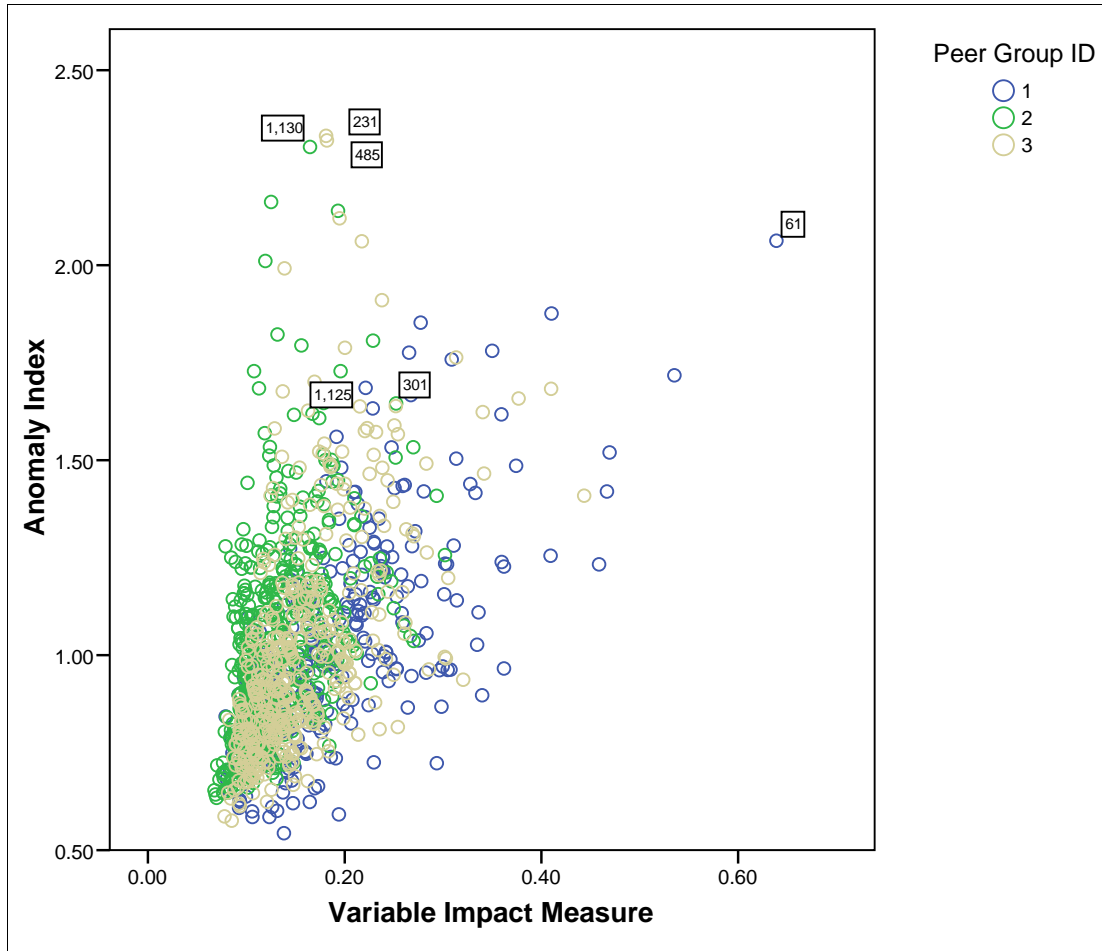


Figure 5. 4 Scatterplot CoSI: Illustrating outliers through respondents' anomaly indices.

Table 5. 4 CoSI: Unusual cases' anomaly indices and Mahalanobis scores.

Respondent	Anomaly Index	Variable Impact	Mahalanobis D <sup>2</sup>	p1	p2
1125	1.62	.17	72.474	.000	.000
301	1.65	.25	70.012	.000	.000
1130	2.303	.165	53.604	.000	.000
231	2.320	.182	51.443	.000	.000
485	2.331	.181	40.945	.002	.000
61	2.063	.639	-	-	-

\* This case was not identified as an outlier by AMOS' Mahalanobis test.

### 5.2.2 Testing for normality

A critically important assumption associated with SEM (covariance based maximum likelihood and generalised least squares only), is the requirement that the data have a multivariate normal distribution (Hair et al., 2006, Byrne, 2001, West et al., 1995, Schumacker and Lomax, 2004). Typically, SEM methodology employs the estimation of parameters using either maximum likelihood (ML) estimation or normal theory generalised least squares (GLS) estimation, both of which demand that the data be multivariate normal (Byrne, 2001). Investigators have often been criticised for failing to notice violations of the normality assumption when employing SEM methodology using ML estimation (West et al., 1995, Byrne, 2001). Following a review of empirical studies examining non-normality in SEM, West et al. (1995) summarised four key findings:

1. As data become increasingly non-normal, the  $\chi^2$  value derived from ML and GLS estimation becomes excessively large, leading to an assumed poor fit. In addition, Byrne (2001) points out that this situation encourages researchers to seek further modification of their hypothesised model in an effort to attain adequate fit to the data, but that these efforts can lead to inappropriate and nonreplicable modifications to otherwise theoretically adequate models.
2. As sample size decreases and non-normality increases, ML and GLS estimators produce analyses that: fail to converge; result in an improper solution; or, at best, yield  $\chi^2$  values that are inflated.
3. When data are non-normal, fit indices such as the Tucker-Lewis Index (TLI; Tucker and Lewis, 1973) and the comparative fit index (CFI; Bentler, 1990), yield values that are modest underestimates.

4. Non-normality can lead to spuriously low standard errors, with degrees of underestimation ranging from moderate to severe.

Given the above problems resulting from the use of ML and GLS estimators under violations of the multivariate normality assumption, it was important to check that the data captured for all three models (AO, CoSI and PNS) was multivariate normal.

AMOS 7 has a built in test for normality which reports values for both skew (SK) and kurtosis (KU) of each indicator variable present in a model, as well as the multivariate value for the complete model, based on Mardia's coefficient and its critical ratio. In Amos 7 the critical ratio for Mardia's coefficient is equal to Mardia's coefficient divided by its standard error. "Assuming normality in very large samples, each of the critical values shown in the table [...] is an observation on a standard normally distributed random variable" (AMOS 7 Help tool; Discussion of Normality Check Example). "There is no generally accepted cut-off value of multivariate kurtosis that indicates non-normality," (Hancock and Mueller, 2006: 273). However, Hancock and Mueller (2006) point out that the guideline offered through the EQS software program suggests that data associated with a value of Mardia's normalised multivariate kurtosis greater than 3, could produce inaccurate results when used with ML estimation.

Univariate normality (skew and kurtosis) and multivariate kurtosis of the data was assessed for all three measurement models. Table 5. 5 shows the skew (SK) and kurtosis (KU) values for AO; it shows the univariate values for each item as well as the overall multivariate kurtosis value for the AO model. The columns labelled 'c.r.' show the z-scores for the SK and KU values (skewness÷standard error, kurtosis÷standard error). The univariate SK c.r. values ranged from (absolute) .071-10.581, with a mean of 4.120.



At the .05 significance level this falls substantially above the 1.96 critical value (for z-scores). The univariate KU c.r. values ranged from .139-7.073, with a mean of 4.328. At the .05 significance level this too falls substantially above the 1.96 critical value. In practice many use the more lenient critical value of 3 (Savalei and Bentler, 2006), however by this standard the mean c.r. are still too high. This would indicate that many of the items have significant non-normal distributions (Savalei and Bentler, 2006, Hancock and Mueller, 2006). As can be seen in Table 5. 5, Mardia's normalised coefficient (c.r.) is 61.934 which indicates a significant departure from multivariate normality.

Table 5. 6 shows the skew and kurtosis values for the CoSI model. The univariate SK c.r. values ranged from (absolute) .239-20.741, with a mean of 11.510. The univariate KU c.r. values ranged from .484-17.418, with a mean of 4.566. SK and KU fell substantially above the critical value, indicating that many of the items have significant non-normal distributions. Mardia's normalised coefficient (c.r.) was 53.852 which also indicates a significant departure from multivariate normality.

**Table 5. 5 AO Assessment of normality (T1 data)**

Variable	min	max	skew	c.r.	kurtosis	c.r.
T1_AO_01	-1.000	1.000	-.694	-10.431	-.018	-.139
T1_AO_02	-1.000	1.000	-.704	-10.581	-.068	-.511
T1_AO_03	-1.000	1.000	-.273	-4.104	-.745	-5.600
T1_AO_04	-1.000	1.000	-.008	-.121	-.941	-7.073
T1_AO_05	-1.000	1.000	-.005	-.071	-.533	-4.008
T1_AO_06	-1.000	1.000	-.157	-2.367	-.861	-6.471
T1_AO_07	-1.000	1.000	-.262	-3.943	-.697	-5.239
T1_AO_08	-1.000	1.000	-.155	-2.337	-.888	-6.676
T1_AO_09	-1.000	1.000	-.130	-1.959	-.757	-5.692
T1_AO_10	-1.000	1.000	-.482	-7.251	-.411	-3.094
T1_AO_11	-1.000	1.000	-.588	-8.835	-.049	-.366
T1_AO_12	-1.000	1.000	-.089	-1.341	-.616	-4.635
T1_AO_13	-1.000	1.000	-.239	-3.589	-.796	-5.987
T1_AO_14	-1.000	1.000	-.104	-1.559	-.641	-4.824
T1_AO_15	-1.000	1.000	-.220	-3.310	-.613	-4.610
Multivariate					75.938	61.934

**Table 5. 6 CoSI Assessment of normality (T1 data)**

Variable	min	max	skew	c.r.	kurtosis	c.r.
T1_CoSI_01_K	-1.000	1.000	-1.379	-20.741	2.197	16.517
T1_CoSI_02_K	-1.000	1.000	-1.304	-19.610	2.316	17.418
T1_CoSI_03_K	-1.000	1.000	-.804	-12.084	.371	2.787
T1_CoSI_04_K	-1.000	1.000	-.746	-11.213	.064	.484
T1_CoSI_05_P	-1.000	1.000	-.901	-13.557	.198	1.486
T1_CoSI_06_P	-1.000	1.000	-1.026	-15.426	.465	3.494
T1_CoSI_07_P	-1.000	1.000	-.878	-13.203	.358	2.691
T1_CoSI_08_P	-1.000	1.000	-1.032	-15.521	.798	5.997
T1_CoSI_09_P	-1.000	1.000	-.724	-10.884	-.178	-1.338
T1_CoSI_10_P	-1.000	1.000	-.369	-5.552	-.764	-5.744
T1_CoSI_11_P	-1.000	1.000	-.779	-11.711	.131	.985
T1_CoSI_12_C	-1.000	1.000	-.804	-12.087	.552	4.149
T1_CoSI_13_C	-1.000	1.000	-.694	-10.443	.122	.916
T1_CoSI_14_C	-1.000	1.000	-.459	-6.903	-.144	-1.080
T1_CoSI_15_C	-1.000	1.000	-.982	-14.773	.539	4.052
T1_CoSI_16_C	-1.000	1.000	-.312	-4.691	-.635	-4.771
T1_CoSI_17_C	-1.000	1.000	-.568	-8.536	-.204	-1.537
T1_CoSI_18_C	-1.000	1.000	.016	.239	-.897	-6.745
Multivariate					78.453	53.852

Table 5. 7 shows the skew and kurtosis values for the PNS model. The univariate SK c.r. values ranged from (absolute) .784-8.465, with a mean of 4.498. The univariate KU c.r. values ranged from 2.603-8.808, with a mean of 6.470. SK and KU fell substantially above the critical value, indicating that several of the items have significant non-normal distributions. Mardia’s normalised coefficient (c.r.) was 18.182 and, whilst lower than

those for the other models, still indicated a highly significant departure from multivariate normality.

**Table 5. 7 PNS Assessment of normality (T1 data)**

Variable	min	max	skew	c.r.	kurtosis	c.r.
T1_PNS_01	-1.000	1.000	-.393	-5.917	-.905	-6.810
T1_PNS_02	-1.000	1.000	.435	6.546	-.911	-6.856
T1_PNS_03	-1.000	1.000	-.562	-8.465	-.346	-2.603
T1_PNS_04	-1.000	1.000	-.252	-3.792	-1.120	-8.434
T1_PNS_05	-1.000	1.000	.175	2.642	-.924	-6.953
T1_PNS_06	-1.000	1.000	-.111	-1.672	-.972	-7.320
T1_PNS_07	-1.000	1.000	-.176	-2.647	-1.175	-8.848
T1_PNS_08	-1.000	1.000	.323	4.856	-1.001	-7.538
T1_PNS_09	-1.000	1.000	.052	.784	-.890	-6.698
T1_PNS_10	-1.000	1.000	.420	6.326	-.590	-4.439
T1_PNS_11	-1.000	1.000	-.387	-5.827	-.620	-4.668
Multivariate					16.675	18.182

Given the non-normal distributions of the data, the SEM implications were considered carefully. Whilst some have argued that the ML estimation can be robust to violations of the normality assumption under certain conditions (Ogasawara, 2003, Browne and Shapiro, 1988, Amemiya and Anderson, 1990, Mooijaart and Bentler, 1991, Satorra, 2001, Savalei, 2008), these conditions were not met in this case, given the severe degree of multivariate non-normality. With this data, both ML and GLS would have been inappropriate to use for model estimation.

AMOS 7 offers two avenues for analysis in dealing with non-normally distributed data when assessing model fit: (a) the Bollen-Stine Bootstrap; and (b) alternative estimation methods such as ULS and ADF.

Bollen and Stine (1993) provided a means to test the null hypothesis that the specified model is correct using the bootstrap process<sup>2</sup>. In other words, the Bollen-Stine Bootstrap provides a probability value associated with the  $\chi^2$ , with the model being rejected if the p-value is  $<.05$ . However, Byrne (2001) points out that like  $\chi^2$ , Bollen-Stine's p-value is very affected by a large sample size (i.e. the test is over-powered), and the researcher is advised to use other measures of fit when this is the case. Given that "large samples can be considered as consisting of more than 500 respondents" (Hair et al., 2006: 748), and the sample consisted of 1,363 respondents, the Bollen-Stine Bootstrap was not used here.

When considering the alternative estimation methods available, the asymptotic distribution free (ADF)<sup>3</sup> estimation was the most appropriate, and has been said to be the most recommended when data is continuous<sup>4</sup> and non-normal (Browne, 1984). It has been advocated by many as a method for dealing with non-normal data (Kline, 2005, Byrne, 1995, Schermelleh-Engel et al., 2003, Schumacker and Lomax, 2004).

There were several advantages to its use:

- It does not assume multivariate normality of the measured variables (Browne, 1984);
- It produces unbiased estimates of the  $\chi^2$  goodness-of-fit test, parameter estimates, and standard errors, which are major theoretical advantages relative to the normal theory-based ML and GLS estimators (West et al., 1995);

---

<sup>2</sup> "Bootstrapping serves as a resampling procedure by which the original sample is considered to represent the population. Multiple subsamples of the same size as the parent sample are then drawn randomly, *with replacement*, from this population and provide the data for empirical investigation of the variability of parameter estimates and fit," (Byrne, 2001: 268-269).

<sup>3</sup> Also known as weighted least squares (WLS) (Bollen, 1989).

<sup>4</sup> Although Likert-type scales are technically ordinal scales, most researchers treat them as continuous variables and use normal theory statistics with them. When there are five or more categories there is relatively little harm in doing this (Johnson and Creech, 1983, Zumbo and Zimmerman, 1993). Out of the 779 Individualised Rating-Scales defined, only six respondents defined rating-scales with fewer than 5 categories (i.e. three or four).

- It is easily applied through AMOS 7;
- Unlike some distribution-free estimators (e.g. unweighted least squares), ADF estimation still provides all key model fit statistics.

The most commonly cited limitation of the ADF estimator is the need for a large sample size, in order to produce stable estimates (West et al., 1995, Schermelleh-Engel et al., 2003, Byrne, 1995). Some have advised that sample sizes of 1000 appear to be necessary with relatively simple models under non-normality (Curran et al., 1994). Given the present sample consisted of 1,363 respondents, it met the criteria. Moreover, split into its two groups for a multi-group comparison of rating-scale used (IRSP: 779 and Likert: 584), the samples met guidelines specified by others where, for example, sample size should approximate  $1.5(p + q)(p + q + 1)$ , where  $p$  equals the number of exogenous variables and  $q$  equals the number of endogenous variables in a model (Joreskog & Sorbom, 1996, as cited in Hancock and Mueller, 2006). Sample size was sufficient to enable the use of the ADF estimator, as long as the measurement models were examined individually (AO, PNS and CoSI) and not bundled together to form a larger measurement model<sup>5</sup>.

### 5.2.3 Assessing measurement model fit

Whilst researchers have often used a rule-of-thumb when considering model acceptability, where fit indices should generally be above .90, it has been shown not to work well with various types of indices, sample sizes, estimators, and distributions (Hu and Bentler, 1995). Hair et al. (2006) offer several general guidelines for determining

---

<sup>5</sup> AO model required an approximate sample size of  $1.5(1+15)(1+15+1) = 408$ . PNS model required an approximate sample size of  $1.5(2+11)(2+11+1) = 273$ . CoSI model required an approximate sample size of  $1.5(3+18)(3+18+1) = 693$ , which is higher than the number of respondents in the Likert group, but was not considered to be high enough to pose a problem based on general guidelines (Hu and Bentler, 1995).

the acceptability of fit for a given measurement model. First, they recommend that multiple fit indices be reported, ideally three or four. Offering an example, they state that “reporting the  $\chi^2$  value and degrees of freedom, the CFI, and the RMSEA will often provide sufficient and unique information to evaluate a model” (Hair et al., 2006: 752). Table 5. 8 illustrates the cut-off values for the most commonly reported model fit indices, together with the source for the guideline. The cut-off values were taken from guidelines that took into consideration sample size and specific model characteristics that were applicable here. For descriptions of each fit index please refer to Hair et al. (2006), or to Schermelleh-Engel et al. (2003).

**Table 5. 8 Guidelines for Goodness of Fit Statistics for large samples.**

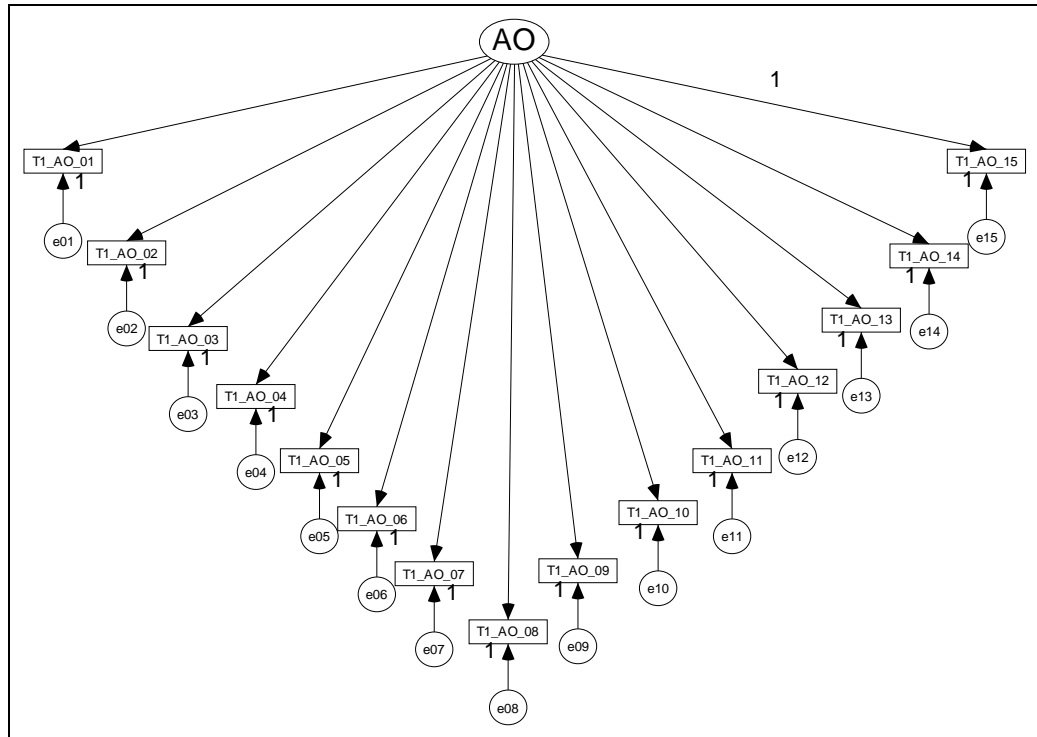
<b>Fit Index</b>	<b>Guideline</b>
$\chi^2$	Where $m < 12$ : Insignificant p-values ( $>.05$ ) with good fit. Where $12 < m < 30$ : Significant p-values can be expected (Type I error rate: $<.05$ )
$\chi^2/df$ ratio	$0 < \chi^2/df \leq 2$ (good fit) $2 < \chi^2/df \leq 3$ (acceptable fit)
CFI	Values $\geq .95$ (good fit) Values $\geq .90$ (acceptable fit)
GFI	Values $\geq .95$ (good fit) Values $\geq .90$ (acceptable fit)
AGFI	Values $\geq .90$ (good fit) Values $\geq .85$ (acceptable fit)
RMSEA	Values $< .05$ , with $p \leq .10$ (good fit) Values $< .07$ , with $p \leq .10$ (acceptable fit)
SRMR	Values $< .05$ (good fit) Values $< .10$ (acceptable fit)

*m* = number of observed variables

Sources for guidelines: (Hair et al., 2006, Schermelleh-Engel et al., 2003, Kline, 2005, Schumacker and Lomax, 2004, Byrne, 2001)

*5.2.3.1 Affective Orientation (AO)*

The Affective Orientation measurement model (Figure 5. 5), with fifteen items, was defined by Booth-Butterfield and Booth-Butterfield (1996).



**Figure 5.5 AO Original Measurement Model.**

The model fit statistics for the AO measurement model are presented for both groups (IRSP and LTRS) in Table 5. 9. Both groups had relatively high  $\chi^2$  values, with the IRSP group producing the highest. However, given that the  $\chi^2$  value is highly sensitive to sample size (Hair et al., 2006, West et al., 1995, Byrne, 2001), and the IRSP group is larger by approximately 200 respondents, this result was unsurprising. In addition, a significant p-value is expected when sample sizes are large, and constructs have more than twelve items (Hair et al., 2006). Particularly worthy of note, is that “sample size has a substantial effect on the  $\chi^2$  statistic based on the ADF estimation” (Hu and Bentler, 1995: 96). Given the large sample size and that the ADF estimation method was used to overcome problems with multivariate non-normality, the implications are that the  $\chi^2$  and its associated p-value are unreliable indications of model fit for both groups. Additionally, although CFI has been reported, it (along with other fit indices such as the TLI) has been found to underestimate model fit for data that is non-normally

distributed (West et al., 1995) and it typically yields “lower values than the threshold level generally perceived as “acceptable” for other normed indices of fit,” (Byrne, 2001: 82). More importantly, CFI, based on ADF estimation, has been shown to over-reject true models (Hu and Bentler, 1995). It is unsurprising, therefore, that both groups produced very low CFI scores, well below the cut-off. The CFI values cannot be relied upon to produce a reliable measure of fit under these conditions. The other fit indices have not been shown to be problematic under ADF estimation. The GFI, in fact, has been shown to be extremely reliable under ADF, when sample sizes are larger than 500 (Hu and Bentler, 1995).

**Table 5.9 AO: Fit Statistics for IRSP and LTRS groups in T1.**

<b>Fit Index</b>	<b>Guideline</b>	<b>IRSP</b>	<b>LTRS</b>
N	-	774	583
$\chi^2$ (df) p	Significant p-values can be expected (Type I error rate: <.05)	400.086 (90) .000	342.036 (90) .000
$\chi^2$ /df ratio	0 < $\chi^2$ /df ≤ 2 (good fit) 2 < $\chi^2$ /df ≤ 3 (acceptable fit)	4.445	3.800
CFI	Values ≥ .95 (good fit) Values ≥ .90 (acceptable fit)	.667	.648
GFI	Values ≥ .95 (good fit) Values ≥ .90 (acceptable fit)	.841	.881
AGFI	Values ≥ .90 (good fit) Values ≥ .85 (acceptable fit)	.789	.841
RMSEA (C.I.)	Values < .05, with p ≤ .10 (good) Values < .07, with p ≤ .10 (acceptable)	.067 (.060-.074)	.069 (.062-.077)
SRMR	Values < .05 (good fit) Values < .10 (acceptable fit)	.153	.153

C.I. = Confidence Interval. N = Sample size.

Bearing all this in mind, it would seem that both groups still did not possess adequate model fit. The IRSP and LTRS GFI indices were .841 and .881 respectively (falling below the .90 cut-off); the AGFI indices were .789 and .841 respectively (falling below the .85 cut-off); the SRMR indices were .153 for both groups (falling above the .10 cut-off). Whilst the IRSP and LTRS RMSEA indices were .067 and .069 respectively (with



both confidence intervals below .08), indicating acceptable fit, this too could have been better.

It was important to examine not only the fit of the model, but also the construct validity, even for a pre-established scale such as the Affective Orientation scale (Hair et al., 2006).<sup>6</sup> Given that the AO measurement model possesses only one latent construct, discriminant validity was not relevant. However, construct validity was assessed by looking at the model's convergent validity, specifically its; factor loadings, average variance extracted and reliability (Hair et al., 2006). The standardised factor loadings were examined for both groups (Table 5. 10), in order to assess the items' communality, using the general rule of thumb outlined by Hair et al. (2006): standardised loading estimates should be .5 or higher, and ideally .7 or higher. Their magnitude, direction and statistical significance were also examined.

---

<sup>6</sup> "The measurement model provides an assessment of convergent and discriminant validity, and the structural model provides an assessment of nomological validity," (Schumacker and Lomax, 2004). Given that the AO *measurement* model is what is being tested, there is no need to check nomological validity.

**Table 5. 10 AO: Standardised factor loadings for T1 data.**

AO Standardised Factor Loadings Time Period 1 Data			
Item regression		IRSP	LTRS
T1_AO_01	<---AO	<b>.820</b>	.788
T1_AO_02	<---AO	.776	<b>.817</b>
T1_AO_03	<---AO	.869	<b>.872</b>
T1_AO_04	<---AO	.705	<b>.772</b>
T1_AO_05	<---AO	<b>.718</b>	.634
T1_AO_06	<---AO	.772	<b>.783</b>
T1_AO_07	<---AO	.848	<b>.855</b>
T1_AO_08	<---AO	.803	<b>.844</b>
T1_AO_09	<---AO	<b>.837</b>	.834
T1_AO_10	<---AO	<b>.809</b>	.738
T1_AO_11	<---AO	.631	<b>.668</b>
T1_AO_12	<---AO	.462	<b>.673</b>
T1_AO_13	<---AO	<b>.899</b>	.888
T1_AO_14	<---AO	.728	<b>.745</b>
T1_AO_15	<---AO	<b>.756</b>	.720

All the standardised factor loadings are above .50, except for item 12 for the IRSP group. Whilst the others are all above .50, some are slightly lower than is ideal. The figures in Table 5. 10 have been emboldened to show in which of the two groups the item made a more substantial contribution to the latent construct. It was interesting to see that six out of the fifteen items had higher standardised loadings in the IRSP group. Two out of the fifteen standardised factor loadings fell below the recommended .7 value for the IRSP group; items 12 (.462) and 11 (.631). Three out of the fifteen standardised factor loadings fell below the ideal .7 value for the LTRS group, namely; items 12 (.673), 11 (.668) and 5 (.634).

It was also important to examine the *construct validity* of the concepts being measured. Researchers have often defined construct validity as “the extent to which an operationalization measures the concept it is supposed to measure” (Bagozzi, 1994: 20). As such, the Average percentage of Variance Extracted (AVE) was examined, as a summary indicator of convergence among a set of construct items (Hair et al., 2006).

Hair et al. (2006) asserted that a good rule of thumb is that an AVE of .5 or higher suggests adequate convergence. The AVE for the Affective Orientation model was .592 for the IRSP group and .607 for the LTRS group, suggesting that on average, less error remains in the items than variance explained by the latent factor structure imposed on the measure. A test of discriminant validity was not necessary with the AO model given it consists only of one factor.

Finally, reliability was examined, as this is also an indicator of convergent validity. As Cronbach's coefficient alpha may underestimate reliability (Hair et al., 2006), the construct reliability (CR) was computed by dividing the squared sum of the standardised regression weights, by this sum added to the sum of indicator measurement error (CR). High construct reliability indicates that internal consistency is present, meaning that the measures all consistently represent the same latent construct. The CR computed was .955 for the IRSP group and .958 for the LTRS group. This indicated high construct reliability, which in turn suggests that high internal consistency existed for both groups. Thus, it seemed as though the measures all represented the same latent construct. Taken together, the data supported the convergent validity of the measurement model in both groups.

Overall, the evidence suggested that whilst convergent validity was present, the original measurement model did not fit well enough for both groups. A number of model diagnostics were examined to assess the model more closely (factor loadings, standardised residuals and modifications indices). They highlighted specific problems areas and indicated how the model might be corrected.

*Modifying the AO Measurement Model*

The fit indices indicated that the data was not fitting the AO measurement model well for either group (Table 5. 9). As already observed, a few of the path estimates were a bit low (for both groups), falling below the ideal .7 value for standardised factor loadings. This indicated potential problems with some of the items. When examining the path estimates for the IRSP data compared to those from the LTRS data (Table 5. 9), there was clearly an overlap for potential problem items. The two lowest item standardised factor loadings for the IRSP data, were AO\_12 (.462) and AO\_11 (.631). For the Likert data, they were AO\_05 (.634) and AO\_11 (.668), with AO\_12 (.673) coming a close third. Items AO\_12 and AO\_11 were loading poorly in both groups.

When examining the standardised residuals, item AO\_12 was scoring above 4.0 in some of the pairings for both IRSP and LTRS groups, which confirmed an unacceptable degree of error (Hair et al., 2006), with item AO\_08 proving very problematic, particularly for the LTRS group.

The modification indices, particularly the error covariances, indicated problems with items AO\_12 and AO\_08 (Table 5. 11, Table 5. 12) showing likely covariance is occurring between the items and/or the item error variances. However, it is not sound practice to specify relationships between them unless there is a theoretical justification for doing so. Given this is a pre-validated scale where observed variables are assumed to be independent, and there is no such theoretical argument for specifying paths

between the items, the recommended approach is to remove these problem items (Byrne, 2001).<sup>7</sup>

On consideration of all the information, it was decided that item AO\_08 be removed first, before re-examining the fit statistics and estimates again to monitor changes (and in particular to re-examine item AO\_12).

**Table 5. 11 AO: IRSP, extract from the Modification Indices covariances table in AMOS.**

IRSP	M.I.	Par Change
e08 <--> e03	5.549	.007
e08 <--> e07	5.096	.008
e09 <--> e08	18.756	-.016
e12 <--> e04	5.341	.015
e12 <--> e06	7.663	-.015
e12 <--> e11	10.730	.016

**Table 5. 12 AO: LTRS, extract from the Modification Indices covariances table in AMOS.**

LTRS	M.I.	Par Change
e08 <--> e01	13.230	-.013
e08 <--> e06	10.766	.015
e08 <--> e07	11.981	.013
e14 <--> e12	8.378	-.014
e15 <--> e12	13.084	.016

Table 5. 13 shows the new fit statistics for the AO model after removing item AO\_08. GFI, AGFI and SRMR fit indices improved for both groups. RMSEA improved only for the LTRS group, but was still an acceptable fit for both.

---

<sup>7</sup> It is worth mentioning that if the items were formative indicators (as opposed to reflective) this approach would not have been employed.

**Table 5. 13 AO (minus item 8): Fit Statistics for IRSP and LTRS groups in T1.**

Fit Index Model	IRSP		LTRS	
	Original	New	Original	New
N	774	774	583	583
$\chi^2$ (df)	400.086 (90)	353.209 (77)	342.036 (90)	276.003 (77)
p	.000	.000	.000	.000
$\chi^2$ /df ratio	4.445	4.587	3.800	3.584
GFI	.841	.855	.881	.900
AGFI	.789	.802	.841	.863
RMSEA (C.I.)	.067 (.060-.074)	.068 (.061-.075)	.069 (.062-.077)	.067 (.058-.075)
SRMR	.153	.131	.153	.127

C.I. = Confidence Interval. N = Sample size.

However, it was still clear that there was further room for improvement, on examination of the standardised regression weights (Table 5. 14), and on inspection of the modification indices and standardised residuals.

**Table 5. 14 AO (minus item 8): Standardised factor loadings for T1 data.**

AO Standardised Factor Loadings Time Period 1 Data			
Item regression		IRSP	LTRS
T1_AO_01	<---AO	<b>.823</b>	.784
T1_AO_02	<---AO	.778	<b>.804</b>
T1_AO_03	<---AO	.860	.860
T1_AO_04	<---AO	.693	<b>.726</b>
T1_AO_05	<---AO	<b>.708</b>	.644
T1_AO_06	<---AO	<b>.733</b>	.706
T1_AO_07	<---AO	<b>.844</b>	.834
T1_AO_08	<---AO	removed	removed
T1_AO_09	<---AO	.825	<b>.827</b>
T1_AO_10	<---AO	<b>.813</b>	.753
T1_AO_11	<---AO	.636	<b>.703</b>
T1_AO_12	<---AO	.470	<b>.713</b>
T1_AO_13	<---AO	<b>.894</b>	.869
T1_AO_14	<---AO	.710	<b>.759</b>
T1_AO_15	<---AO	.745	<b>.758</b>

Item AO\_12 was still loading poorly for the IRSP group. In addition, both items AO\_12 and AO\_06 were performing poorly on the modification indices in both groups, particularly item AO\_06 (see Table 5. 15 and Table 5. 16).

**Table 5. 15 AO (minus item 8): IRSP, extract from the Modification Indices covariances table.**

<b>IRSP</b>	<b>M.I.</b>	<b>Par Change</b>
e06 <--> e04	17.316	.031
e12 <--> e06	10.154	-.019
e12 <--> e11	9.573	.016
e13 <--> e06	6.389	-.010

**Table 5. 16 AO (minus item 8): LTRS, extract from the Modification Indices covariances table.**

<b>LTRS</b>	<b>M.I.</b>	<b>Par Change</b>
e06 <--> e02	8.749	-.013
e06 <--> e03	15.479	.018
e06 <--> e04	21.558	.029
e06 <--> e05	8.947	-.015
e14 <--> e12	6.806	-.013
e15 <--> e12	9.218	.014

On consideration of all the information, it was decided that item AO\_06 should be removed. This process of re-examining the fit statistics, modification indices, path estimates and standardised residuals, was repeated with inappropriately performing items removed. This continued until an acceptable level of fit was achieved for the remaining items, with the unacceptable level of error removed. In total, six of the fifteen items were removed: AO\_08, AO\_06, AO\_11, AO\_12, AO\_04 and AO\_14. It is worthy of note that items AO\_11 and AO\_14 were mentioned, in section 4.5.3.4, as items that were consistently flagged as a problem by the MBA respondents in the pilot study. It is therefore appropriate that they were removed through statistical inspection of the data. Figure 5. 6 shows the resultant standardised factor loadings for both the IRSP and the LTRS data.

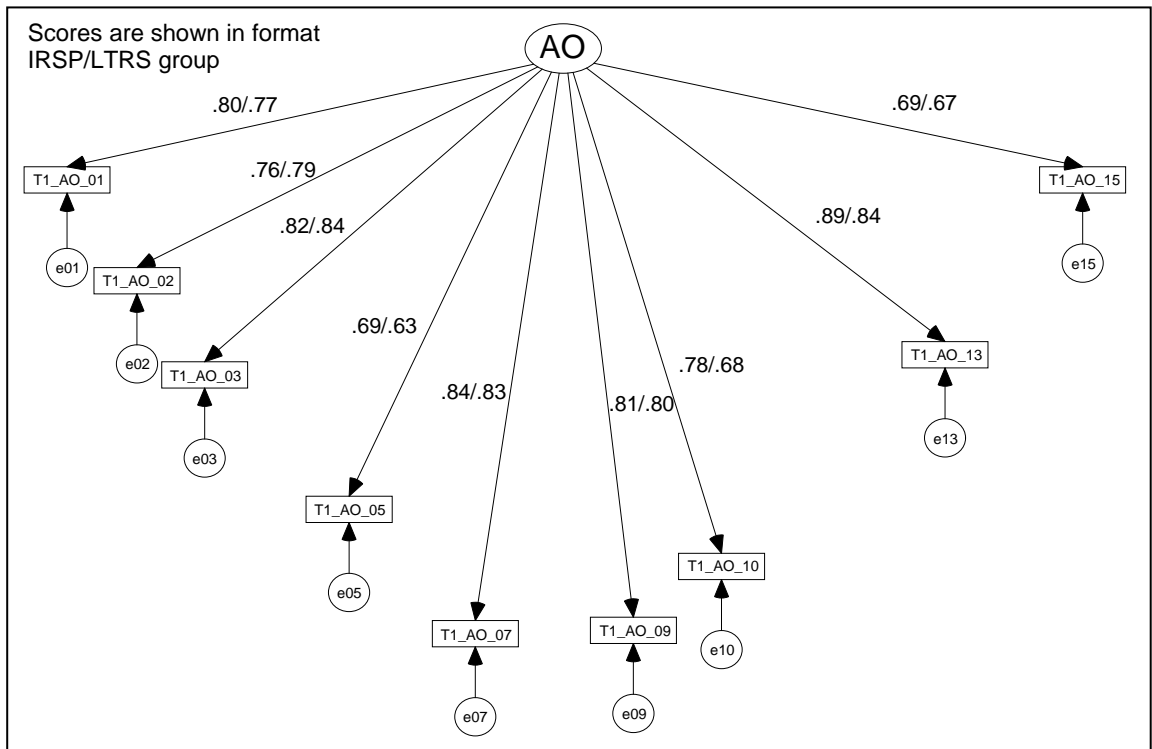


Figure 5. 6 Re-specified AO model: Standardised regression weights shown for both groups in T1.

Table 5. 17 compares the fit statistics of the original AO model with the re-specified AO model. From the figures one can see that the model fit improved significantly as a result of the removal of problematic items. Whilst the p-value remained significant for Type I error, this is not unusual for groups with large sample sizes, and the other fit statistics show good model fit for both groups, although slightly better for the LTRS group. The AVE for the re-specified AO model was .623 for the IRSP group and .585 for the LTRS group. Whilst the AVE went up slightly for the IRSP group, and went down slightly for the LTRS, both are still acceptable, which meant that on average, there was still less error in the items than variance explained by the latent factor structure. The CR computed for the re-specified model was .937 for the IRSP group and .926 for the LTRS group, a slight reduction for both groups. These figures, however, were still very high, suggesting that high internal consistency still existed for both



groups. Thus, it seemed as though the measures all represented the same latent construct for the re-specified measurement model, supporting its convergent validity in both groups. This re-specified model was carried forward to the next stage of the analyses.

**Table 5. 17 Original vs Re-specified AO model: Fit Statistics for IRSP and LTRS groups in T1.**

Model	IRSP		LTRS	
	Original	Re-specified	Original	Re-specified
N	774	774	583	583
$\chi^2$ (df)	400.086 (90)	112.962 (27)	342.036 (90)	90.231 (27)
p	.000	.000	.000	.000
$\chi^2$ /df ratio	4.445	4.184	3.800	3.342
GFI	.841	.937	.881	.954
AGFI	.789	.895	.841	.923
RMSEA (C.I.)	.067 (.060-.074)	.064 (.052-.077)	.069 (.062-.077)	.063 (.049-.078)
SRMR	.153	.077	.153	.060
AVE	.592	.623	.607	.585
CR	.955	.937	.958	.926

### 5.2.3.2 Personal Need for Structure (PNS)

The Personal Need for Structure (PNS) measurement model (Figure 5. 7) was defined by Neuberg and Newsom (1993). It is a construct consisting of two dimensions (Desire for Structure – DS, and Response to Lack of Structure – RLS), and is thus a two-factor model.

DS = Desire for Structure  
 RLS = Response to Lack of Structure

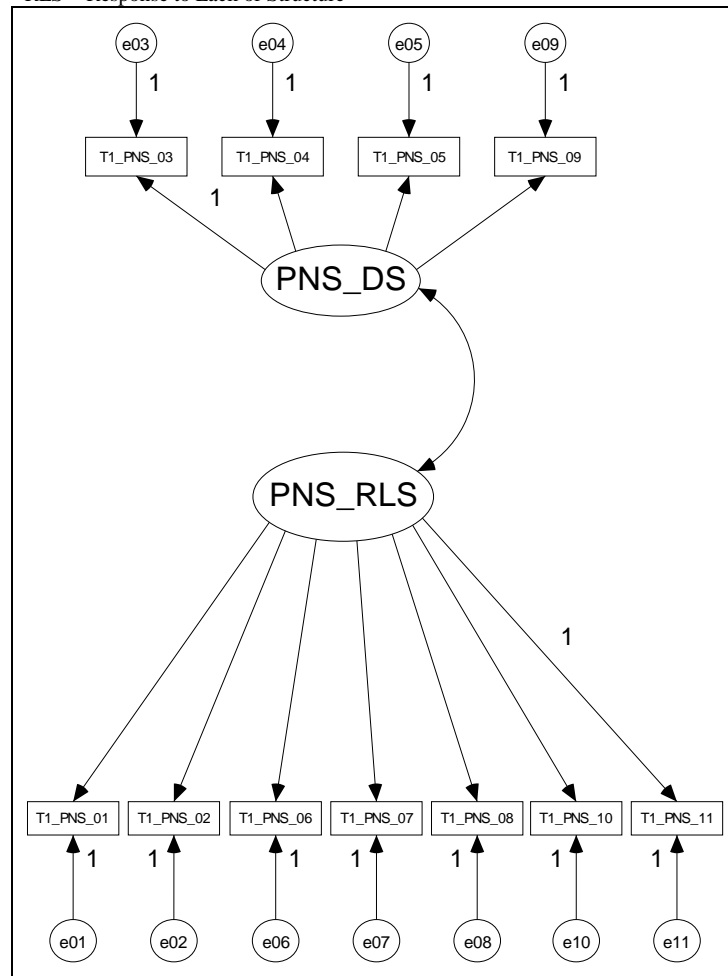


Figure 5. 7 PNS Original Measurement Model.

The model fit statistics for the PNS measurement model are presented for both groups (IRSP and LTRS) in Table 5. 18.

**Table 5. 18 PNS: Fit Statistics for IRSP and LTRS groups in T1.**

Fit Index	Guideline	IRSP	LTRS
N	-	777	583
$\chi^2$ (df) p	Significant p-values can be expected (Type I error rate: <.05)	234.197 (43) .000	206.008 (43) .000
$\chi^2$ /df ratio	0 < $\chi^2$ /df ≤ 2 (good fit) 2 < $\chi^2$ /df ≤ 3 (acceptable fit)	5.446	4.791
GFI	Values ≥ .95 (good fit) Values ≥ .90 (acceptable fit)	.924	.940
AGFI	Values ≥ .90 (good fit) Values ≥ .85 (acceptable fit)	.883	.908
RMSEA (C.I.)	Values < .05, with p ≤ .10 (good) Values < .07, with p ≤ .10 (acceptable)	.076 (.066-.085)	.081 (.070-.092)
SRMR	Values < .05 (good fit) Values < .10 (acceptable fit)	.0832	.0739

C.I. = Confidence Interval. N = Sample size.

The  $\chi^2$  values and their associated p-values have been reported for information purposes, even though they have been shown to be an unreliable indication of model fit in this context due to the sample size and ADF estimator, as already mentioned. The IRSP group produced the higher  $\chi^2$ /df ratio of the two groups at 5.446. Both the IRSP and LTRS groups demonstrated adequate fit across several of the indices, with the LTRS group performing marginally better: GFI indices were .924 and .940 respectively; AGFI indices were .883 and .908; and the SRMR indices were .0832 and .0739. Whilst the IRSP and LTRS RMSEA indices were .076 and .081 respectively, indicating adequate fit, the upper end of the confidence interval was on the high side for both groups (although worse for LTRS); both were above the .08 cut-off at .085 and .092 respectively. In summary, the fit indices suggested acceptable fit for both groups, although the RMSEA indicated that there may be room for improvement.

A brief check on the construct validity of the PNS model was conducted. To this aim, the model's convergent and discriminant validity was assessed. Firstly, its factor loadings (Table 5. 19), variance extracted and reliability (Table 5. 20) were examined to assess convergent validity.

**Table 5. 19 PNS: Standardised factor loadings for T1 data.**

<b>PNS Standardised Factor Loadings Time Period 1 Data</b>			
Item regression		IRSP	LTRS
T1_PNS_03	<--- PNS_DS	0.727	<b>0.763</b>
T1_PNS_04	<--- PNS_DS	0.551	<b>0.615</b>
T1_PNS_05	<--- PNS_DS	<b>0.621</b>	0.581
T1_PNS_09	<--- PNS_DS	0.783	<b>0.806</b>
T1_PNS_11	<--- PNS_RLS	0.616	<b>0.687</b>
T1_PNS_10	<--- PNS_RLS	<b>0.678</b>	0.578
T1_PNS_08	<--- PNS_RLS	<b>0.653</b>	0.597
T1_PNS_07	<--- PNS_RLS	0.720	<b>0.723</b>
T1_PNS_06	<--- PNS_RLS	0.747	<b>0.794</b>
T1_PNS_02	<--- PNS_RLS	<b>0.588</b>	0.527
T1_PNS_01	<--- PNS_RLS	0.743	<b>0.748</b>

**Table 5. 20 PNS Original Model: Factors' Average Variance Extracted and Construct Reliability.**

<b>Group</b>	<b>IRSP</b>		<b>LTRS</b>	
	<b>DS</b>	<b>RLS</b>	<b>DS</b>	<b>RLS</b>
AVE	.458	.463	.487	.450
CR	.768	.857	.788	.849

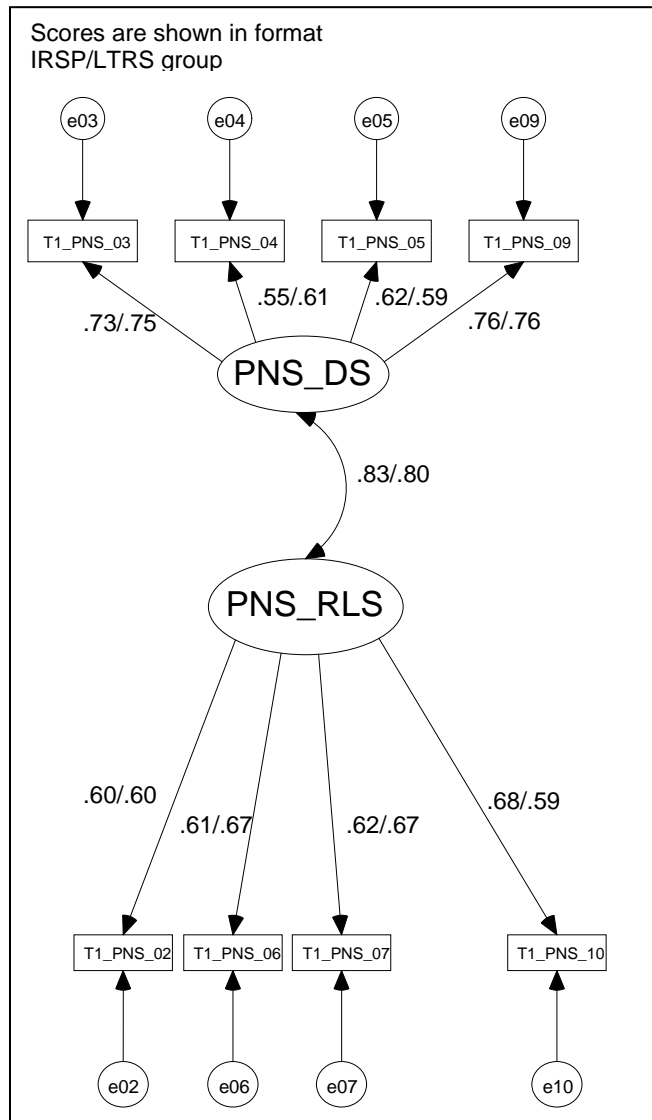
The standardised factor loadings, shown in Table 5. 19, are all above the recommended .50 cut-off (Hair et al., 2006), however, several of them are lower than is ideal. Four out of the eleven items from the LTRS, and two from the IRSP, are below .60. Four of the items had higher loadings in the IRSP group, shown in bold. The average variance extracted (AVE) and construct reliability (CR) was calculated for each of the two latent constructs (factors), in both groups. Although in both groups the reliabilities are above the .70 cut-off point (suggesting adequate internal consistency, especially for the RLS latent variable), the AVE scores are marginally too low in all cases (Table 5. 20). This would normally suggest that more error remained in the items than variance explained by the latent factor structure imposed on the measures.

Next, the discriminant validity was assessed as it was necessary to check the extent to which the two PNS constructs are distinct from one another. Hair et al. (2006)

recommends that the best test of discriminant validity is to compare the variance extracted measures for any two constructs with the square of the correlation estimate between these two constructs. The former should, ideally, be higher than the latter. The inter-construct correlation between DS and RLS for the IRSP group (as reported by AMOS) was .787. Squared, this equals .619. The squared inter-construct correlation for the IRSP group was higher than both the AVE scores for each construct (Table 5. 20). The squared inter-construct correlation for the LTRS group was  $.727^2 = .529$ , and was also higher than the AVE scores for both constructs. Whilst, this would normally suggest that there is an inadequate level of discriminant validity in the measurement model (for both groups), this is not an area of concern with the PNS scale. The reason being, that this result is in line with those of the original study by Neuberg and Newsom (1993). They highlighted that the original one-factor version of the scale appeared to capture two related, but conceptually distinct elements; the *desire* for structure, and how one *responds* to its absence (behaviourally). They tested the one-factor version against the two-factor version using CFA and found that the two-factor model fit the data better. Additionally, they reported that the two factors correlated highly (with inter-factor correlations ranging from .54 – .75 across six sample groups). They asserted that the two-factor model is preferred, and that one would expect the two factors to be highly related. As such, given this result is in accordance with that of Neuberg and Newsom, it did not require corrective attention.

Standardised factor loadings, modification indices and standardised residuals were examined, with poorly performing items removed. The resultant model is shown in Figure 5. 8, and consisted of the removal of items PNS\_08, PNS\_01 and PNS\_11. Interestingly, both items PNS\_08 and PNS\_01 had been flagged as potential problem

items during the pilot test with the MBA students, so it would seem appropriate that they were removed. The figure shows the standardised regression weights.



**Figure 5. 8 Re-specified PNS model: Standardised regression weights shown for both groups in T1**

Table 5. 21 compares the fit statistics of the original PNS model with the re-specified PNS model. From the figures one can see that the model fit significantly improved as a result of the removal of the problematic items. The fit statistics show good model fit for both groups, with a particularly strong fit with the IRSP group.

**Table 5. 21 Original vs Re-specified PNS model: Fit Statistics for IRSP and LTRS groups in T1.**

Fit Index	IRSP		LTRS	
	Original	Re-specified	Original	Re-specified
N	777	777	583	583
$\chi^2$ (df)	234.197 (43)	42.687 (19)	206.008 (43)	61.178 (19)
p	.000	.001	.000	.000
$\chi^2$ /df ratio	5.446	2.247	4.791	3.220
GFI	.924	.983	.940	.978
AGFI	.883	.968	.908	.958
RMSEA	.076	.040	.081	.062
(C.I.)	(.066-.085)	(.024-.056)	(.070-.092)	(.045-.079)
SRMR	.083	.035	.074	.045

Table 5. 22 shows the AVE and CR figures for the original and re-specified PNS models for both groups. The AVE for both factors went down with both groups, especially for the RLS factor. The CR figures worsened in both groups especially for the RLS factor, although they were higher than the minimum recommended cut-off of .7. However, it was clear that internal consistency, whilst acceptable, was not very strong. Whilst a clear test of discriminant validity could not be conducted, given the two constructs are expected to be highly related, Stage 3 of this chapter further extends the examination of validity for the psychometric constructs. In Stage 3, all the psychometric constructs are examined for validity in relation to each other.

This re-specified model was carried forward to the Stage 2 of the analyses, in order to see how *equivalent* the results were for both groups.

**Table 5. 22 PNS: Factors' AVE and CR for Original Model vs Re-specified Model.**

Group	IRSP		LTRS	
	DS	RLS	DS	RLS
AVE - Original Model	.458	.463	.487	.450
AVE - Re-specified Model	.449	.395	.465	.402
CR - Original Model	.768	.857	.788	.849
CR - Re-specified Model	.763	.722	.774	.728

5.2.3.3 Cognitive Style Indicator (CoSI)

The Cognitive Style Indicator (CoSI) measurement model was defined by Cools and van den Broeck (2007), and is illustrated in Figure 5. 9.

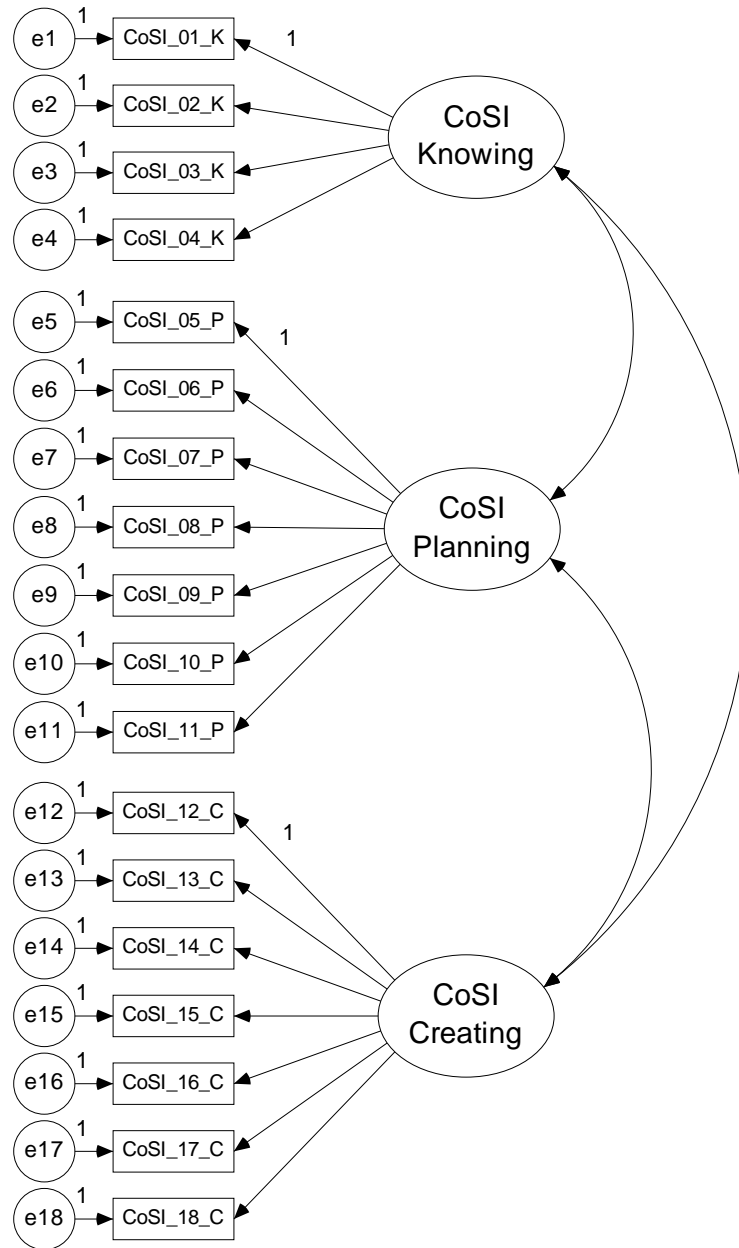


Figure 5. 9 CoSI Original Measurement Model.

The model fit statistics for the CoSI measurement model with three sub-dimensions are presented for both groups (IRSP and LTRS) in Table 5. 23.



**Table 5. 23 CoSI: Fit Statistics for IRSP and LTRS groups in T1.**

Fit Index	Guideline	IRSP	LTRS
N	-	775	582
$\chi^2$ (df) p	Significant p-values can be expected (Type I error rate: <.05)	527.224 (132) .000	558.607 (132) .000
$\chi^2$ /df ratio	0 < $\chi^2$ /df ≤ 2 (good fit) 2 < $\chi^2$ /df ≤ 3 (acceptable fit)	3.994	4.232
GFI	Values ≥ .95 (good fit) Values ≥ .90 (acceptable fit)	.862	.854
AGFI	Values ≥ .90 (good fit) Values ≥ .85 (acceptable fit)	.821	.811
RMSEA (C.I.)	Values < .05, with p ≤ .10 (good) Values < .07, with p ≤ .10 (acceptable)	.062 (.057-.068)	.075 (.068-.081)
SRMR	Values < .05 (good fit) Values < .10 (acceptable fit)	.083	.135

C.I. = Confidence Interval. N = Sample size.

The LTRS group produced the higher  $\chi^2$ /df ratio of the two groups at 4.232. Both the IRSP and LTRS groups demonstrated inadequate fit across the GFI and AGFI indices, with the IRSP group performing marginally better: GFI indices were .862 and .854 respectively; AGFI indices .821 and .811 respectively. The SRMR index indicated an acceptable fit for the IRSP group at .083, however the LTRS group had a significantly poorer score of .135. Whilst the IRSP and LTRS RMSEA indices were .062 and .075 respectively, indicating adequate fit, the upper end of the confidence interval was on the high side for the LTRS group, at .081. In summary, the indices suggested a better model fit for the IRSP group. However, the fit indices suggested the model could fit better in both groups.

A brief check on the construct validity of the CoSI model was conducted; the model's convergent and discriminant validity was assessed. Firstly, its factor loadings (Table 5. 24), variance extracted and reliability (Table 5. 25) were examined to assess convergent validity. It is worth noting that eleven out of the eighteen items had higher item factor loadings in the IRSP group, shown in bold.

**Table 5. 24 CoSI: Standardised factor loadings for T1 data.**

CoSI Standardised Factor Loadings Time Period 1 Data			
Item regression		IRSP	LTRS
T1_CoSI_01_K	<--- CoSI_Knowing	0.492	<b>0.648</b>
T1_CoSI_02_K	<--- CoSI_Knowing	0.718	<b>0.721</b>
T1_CoSI_03_K	<--- CoSI_Knowing	<b>0.883</b>	0.796
T1_CoSI_04_K	<--- CoSI_Knowing	0.713	<b>0.765</b>
T1_CoSI_05_P	<--- CoSI_Planning	<b>0.761</b>	0.633
T1_CoSI_06_P	<--- CoSI_Planning	<b>0.747</b>	0.500
T1_CoSI_07_P	<--- CoSI_Planning	<b>0.752</b>	0.730
T1_CoSI_08_P	<--- CoSI_Planning	<b>0.659</b>	0.585
T1_CoSI_09_P	<--- CoSI_Planning	<b>0.716</b>	0.559
T1_CoSI_10_P	<--- CoSI_Planning	0.587	<b>0.655</b>
T1_CoSI_11_P	<--- CoSI_Planning	<b>0.695</b>	0.646
T1_CoSI_12_C	<--- CoSI_Creating	0.610	<b>0.746</b>
T1_CoSI_13_C	<--- CoSI_Creating	0.673	<b>0.697</b>
T1_CoSI_14_C	<--- CoSI_Creating	0.740	<b>0.768</b>
T1_CoSI_15_C	<--- CoSI_Creating	<b>0.700</b>	0.617
T1_CoSI_16_C	<--- CoSI_Creating	<b>0.727</b>	0.720
T1_CoSI_17_C	<--- CoSI_Creating	<b>0.801</b>	0.730
T1_CoSI_18_C	<--- CoSI_Creating	<b>0.608</b>	0.462

**Table 5. 25 CoSI Original Model: Factors' Average Variance Extracted and Construct Reliability.**

Group	IRSP			LTRS		
	Creating	Planning	Knowing	Creating	Planning	Knowing
AVE	.486	.497	.511	.468	.384	.540
CR	.868	.873	.801	.858	.811	.823

The standardised factor loadings, shown in Table 5. 24, are all above the recommended .50 cut-off (Hair et al., 2006) except for item CoSI\_01 (for the IRSP group) and CoSI\_18 (for the LTRS group). Four out of the eighteen items from the LTRS, and two from the IRSP, are below .60. The average variance extracted (AVE) and construct reliability (CR) was calculated for each of the three latent constructs (factors), in both groups. Although the reliabilities for all constructs, in both groups, fall well above the .70 cut-off point (suggesting adequate internal consistency), the AVE scores are marginally too low in some cases, namely for the Creating and Planning latent factors (Table 5. 25). The LTRS group had lower AVE scores than the IRSP group on two of the three factors, with LTRS' Planning factor being substantially lower than the

benchmark figure. This suggested that more error remained in the items than variance explained by the latent factor structure imposed on the measures.

Next, the discriminant validity was assessed for the CoSI model. It was necessary to check the extent to which the three constructs are truly distinct from one another. The inter-factor correlations are shown in Table 5. 26.

**Table 5. 26 CoSI: Inter-factor correlations T1.**

<b>Factors</b>	<b>Groups</b>			
	<b>IRSP</b>		<b>LTRS</b>	
	Correlations	Correlations <sup>2</sup>	Correlations	Correlations <sup>2</sup>
Creating <---> Planning	-.074	.005	.371	.138
Planning <---> Knowing	.375	.141	.410	.168
Creating <---> Knowing	.331	.110	.483	.233

The squared inter-construct correlations for the IRSP group (Table 5. 26) were lower than all the AVE scores for each construct (Table 5. 25). The squared inter-construct correlations for the LTRS group were also lower than the AVE scores for each of the constructs. This suggested an adequate level of discriminant validity in the measurement model for both groups.

Standardised factor loadings, modification indices and standardised residuals were examined, and poorly performing items removed. The resultant model is shown in Figure 5. 10, with items CoSI\_14, CoSI\_12, CoSI\_02 and CoSI\_18 removed. The figure shows the standardised regression weights for both groups.

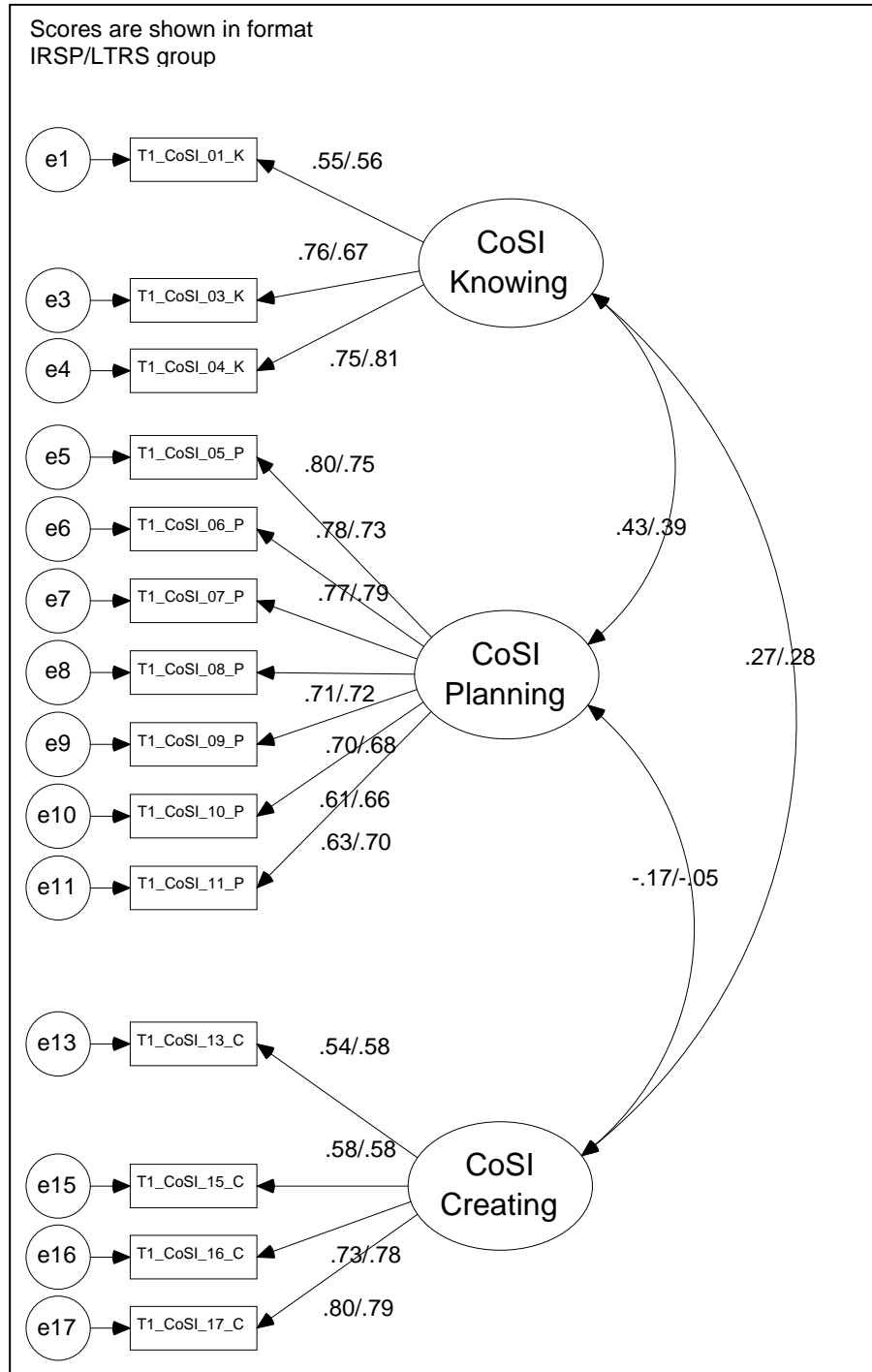


Figure 5. 10 Re-specified CoSI model: Standardised regression weights shown for both groups in T1

Table 5. 27 compares the fit statistics of the original CoSI model with the re-specified CoSI model. From the figures, one can see that the model fit improved as a result of the removal of the problematic items. The fit statistics show an acceptable model fit for

both groups, with a slightly stronger fit with the LTRS group. This re-specified model was carried forward to the next stage of the analyses.

**Table 5. 27 Original vs Re-specified CoSI model: Fit Statistics for IRSP and LTRS groups in T1.**

Fit Index Model	IRSP		LTRS	
	Original	Re-specified	Original	Re-specified
N	775	775	582	582
$\chi^2$ (df)	527.224 (132)	260.189 (74)	558.607 (132)	214.371 (74)
p	.000	.000	.000	.000
$\chi^2$ /df ratio	3.994	3.516	4.232	2.897
GFI	.862	.909	.854	.921
AGFI	.821	.871	.811	.888
RMSEA (C.I.)	.062 (.057-.068)	.057 (.050-.065)	.075 (.068-.081)	.057 (.048-.066)
SRMR	.083	.064	.135	.069

Table 5. 28 shows the AVE and CR figures for the original and re-specified CoSI models. The AVE values for factors ‘Creating’ and ‘Knowing’ went down slightly in the IRSP group, and the AVE for ‘Knowing’ also went down for the LTRS group. AVE values for ‘Planning’ improved in both groups, although most significantly in the LTRS group where it jumped from .384 to .518. Two of the three factors (‘Creating’ and ‘Knowing’) had AVE values marginally below the .5 cut-off in both groups. CR values improved in both groups for the ‘Planning’ factor, although were significantly down for the other two factors. All CR values, however, were still above the recommended .7 cut-off. Despite some of the factors having AVE values that were lower than the recommended cut-off, in all cases (for both groups) they were still significantly larger than the inter-factor covariance estimates shown in Figure 5. 10, suggesting that sufficient discriminant validity was present. This fact, combined with the adequate CR values, meant that the CoSI model continued to be examined in further CFA analysis beyond Stage 2.

**Table 5. 28 CoSI: Factors' AVE and CR for Original Model vs Re-specified Model**

Group Factor	IRSP			LTRS		
	Creating	Planning	Knowing	Creating	Planning	Knowing
AVE - Original Model	.486	.497	.511	.468	.384	.540
AVE - Re-specified Model	.450	.515	.481	.476	.518	.473
CR - Original Model	.868	.873	.801	.858	.811	.823
CR - Re-specified Model	.762	.880	.732	.781	.882	.725

### 5.3 Stage 2: Testing multi-group measurement model equivalence

*Objective:* To compare the IRSP with the original validated LTRS version of the AO, PNS and CoSI.

As the re-specified models for all three constructs indicate acceptable fits for both the IRSP and LTRS groups separately, this would suggest that the IRSP has, at the very least, managed to capture respondent data equally as well as the LTRS. This shows that the measurement models have achieved loose cross-validation across both groups. However, before any further multi-group comparisons could be conducted it was important to establish measurement equivalence between the groups through cross-validation in SEM (Hair et al., 2006, Steenkamp and Baumgartner, 1998, Vandenberg and Lance, 2000). This involved a more rigorous investigation of the degree to which one group produced the same results as the other group, by applying a series of progressively more rigorous tests across the groups. Table 5. 29 summarises the six stages of cross-validation outlined by Hair et al. (2006: 820-821), from least rigorous to most rigorous.

**Table 5. 29 Stages of multiple group cross-validation, as per Hair et al. (2006).**

Type	Description
1. Loose cross-validation	Acceptable model fit in both groups, conducted separately.
2. Equivalent covariance matrices	When two groups have equivalent covariance matrices. As such Hair et al. (2006) state that theoretically this test is redundant, with its usefulness and diagnostic value having been questioned. They advise that researchers proceed straight to the next step.
3. Factor structure equivalence	Sometimes known as configural invariance. This test involves simultaneously estimating CFA models using data from both groups. Only the factor structure is constrained between groups. Here, the fit statistics refer to how well the model fits <i>both</i> covariance matrices.
4. Factor loading equivalence	Constrains the loading estimates to be equal in each group. Here one can examine the new model fit statistics to assess the validity of this model.
5. Factor loading and inter-factor covariance equivalence	This test adds the constraint that the inter-factor covariance terms are equal between samples.
6. Factor loading, inter-factor covariance, and error variance equivalence	This final test is sometimes referred to as tight cross-validation. It adds the constraint that the error variance associated with each residual is equal between groups

In this particular case, it was important to examine whether the rating-scales were being used similarly in both groups. This issue involves *metric equivalence*, a measure of *measurement invariance*, which provides the researcher with an indication as to whether or not people from different groups are interpreting and using rating-scales in the same way. In this context, respondents were using *different* rating-scales in both groups, but it was important to see whether differences between the values obtained could be compared (across both groups). This was necessary to enable meaningful comparisons to be made about the strength of relationships between constructs from one group to the other (Hair et al., 2006). Full metric invariance is present when factor loading equivalence has been established between groups (step 4. in Table 5. 29).

Of additional importance was whether the quantifiable meanings of the rating-scales were the same in both groups; this issue is termed *scalar invariance*. The presence of *scalar equivalence* means that amounts have the same meaning between the two groups being considered. Scalar equivalence exists when the intercept terms for each measured variable are invariant between groups being studied. A test for scalar invariance was not

explicitly listed in Hair et al.'s (2006) six-step cross-validation outline (although it would sit between steps 4. and 5. in terms of rigour). Whilst scalar equivalence has received the least attention out of all the different tests of measurement invariance (Vandenberg and Lance, 2000), it is becoming more popular with researchers, and is indeed an issue for those wishing to compare factor means between groups. When both *metric equivalence* and *scalar invariance* are present between groups, *strong factorial invariance* is said to exist, rendering sound comparisons of means. Unfortunately, it was not possible to conduct a test of scalar equivalence for the three measurement models across groups. This is because AMOS 7 does not compute a multi-group test of invariance of means and/or intercepts, when the ADF estimation method is employed (which was necessary in this case due to the non-normal distribution of the data). Whilst many studies have been published in which multi-group comparisons of means were conducted without assessing either metric or scalar invariance (Childers and Rao, 1992, Durvasula et al., 1993, Dawar and Parker, 1994, Dahlstrom and Nygaard, 1995, Ferrando, 2000), it would have been ideal to be able to evidence both in this study, given that both scalar and metric invariance are needed to make trustworthy valid comparisons (Hair et al., 2006). As a result of this, it could not be claimed with reasonable probability that *strong factorial invariance* existed along with the corresponding ability to make comparisons of means. However, this aim was met in part through the pursuit of metric equivalence.

It is worth adding that neither invariance of factor (co)variances nor invariance of error variances is necessary for comparing means (Horn and McArdle, 1992, Meredith, 1993). This stems from the fact that equal construct reliability is not necessary for mean comparisons (Rock et al. 1998 as cited in Steenkamp and Baumgartner, 1998).



Construct reliability is affected by item loadings, error variances, and construct variances. However, by assumption, errors are independent with an expectation of zero, so they are not expected to affect the latent means. Additionally, there is no conceptual or statistical reason why the construct variances should be equal across groups in order for comparisons of means to be meaningful. This leaves the factor loadings as the only remaining determinant of reliability, and their invariance is incorporated into Steenkamp and Baumgartner's (1998) concept of measurement invariance. In summary, each of the three measurement models were tested in a multi-group design, from steps 3 through to 6 in Table 5. 29 (with steps 5 and 6 being non-essential prerequisites for a comparison of means). Even so, steps 5 and 6 were still included in order to see just how *tight* the cross-validation between the two groups was, for each model. These results are outlined next.

### 5.3.1 AO: Measurement Invariance for IRSP vs LTRS groups

Table 5. 30 recaps the loose cross-validation fit statistics between the groups for the re-specified AO measurement model.

**Table 5. 30 Re-specified AO model: Fit Statistics for IRSP and LTRS groups in T1**

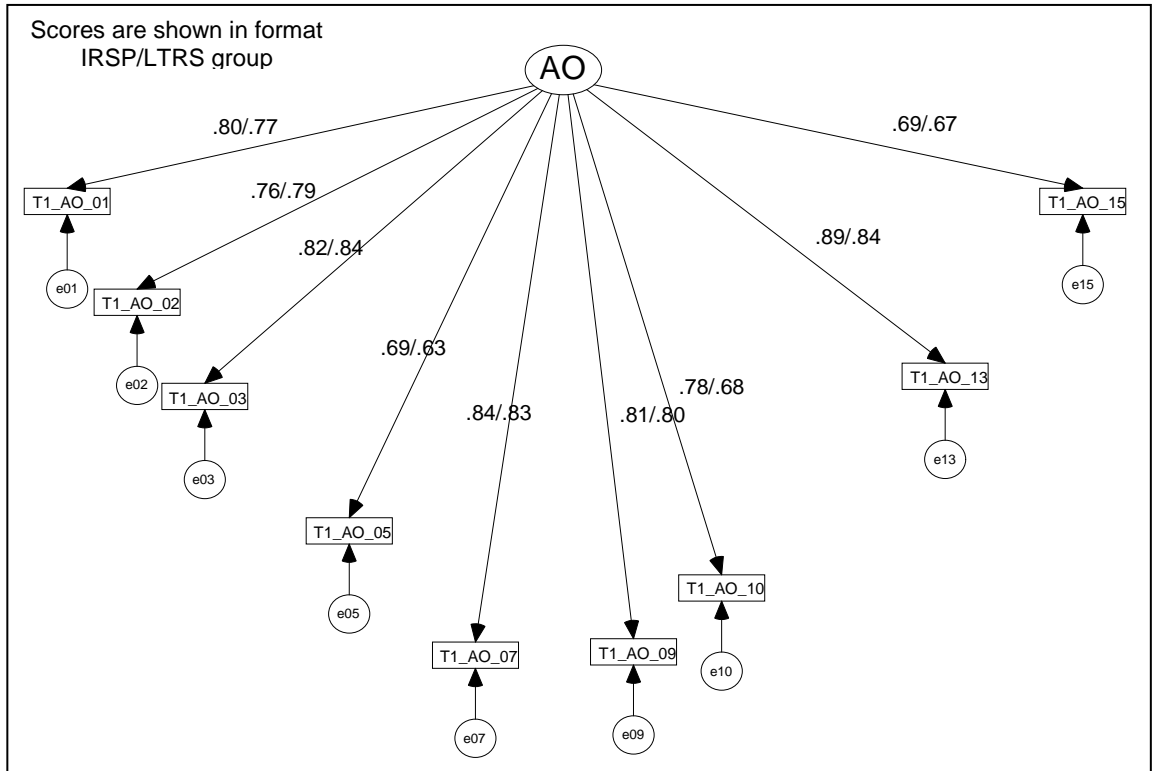
<b>Fit Index</b>	<b>IRSP</b>	<b>LTRS</b>
N	774	583
$\chi^2$ (df)	112.962 (27)	90.231 (27)
p	.000	.000
$\chi^2$ /df ratio	4.184	3.342
GFI	.937	.954
AGFI	.895	.923
RMSEA (C.I.)	.064 (.052-.077)	.063 (.049-.078)
SRMR	.077	.060

#### 5.3.1.1 Factor Structure Equivalence

Next, the AO measurement model was examined in both samples, as in loose cross-validation, but this time the model was estimated in each group simultaneously rather than separately. The resultant fit indices referred to how accurately the measurement

model reproduced the observed covariance matrix for both the IRSP group and the LTRS group. Hair et al. (2006) refer to this as the TF (totally free) model.

As expected, the model  $\chi^2$  for the multi-group CFA (IRSP vs LTRS) equals the value obtained by adding the two  $\chi^2$  values from the loose-validation process together (Table 5. 30). That value was 203.195 with 54 degrees of freedom ( $p = .000$ ), and with a  $\chi^2/df$  of 3.763. As already mentioned, the highly significant  $p$ -value was expected, given the large sample size. The RMSEA for the multi-group model was 0.045 with a 90% confidence interval of .039 to .052, the SRMR was .077, the GFI was .946 and the AGFI was .910. On the whole, these results support the multi-group AO measurement model. Thus, the same factor structure is appropriate in either sample, so factor structure equivalence has been supported. Figure 5. 11 displays the resulting parameter estimates for each group. The parameters estimated generally support the model. Items AO\_05, AO\_10 and AO\_15 were loading somewhat lower than the others, and had lower estimates in the LTRS group (particularly item AO\_10). However, they met the minimum cut-off guidelines, and the model's overall fit was acceptable nonetheless.



**Figure 5. 11 AO Re-specified model: Multi-group factor structure equivalence (TF model) parameter estimates.**

5.3.1.2 *Factor Loading Equivalence and Error Variances Equivalence*

The next test constrains the CFA model so that the factor loading estimates in the two groups are equal; a test for metric equivalence. Factor loading equivalence is tested by examining the effects of adding this constraint on the fit of the TF model. Table 5. 31 displays the fit statistics associated with both models, constraining the factor loading estimates in the IRSP group to equal those in the LTRS group. It also contains results for the other models testing more rigorous degrees of equivalence (steps 4-6 in Table 5. 29).

**Table 5. 31 Cross-Validation: Equivalence models for multi-group AO in T1.**

Model	$\chi^2$	DF	P	$\chi^2/DF$	RMSEA (C.I.)	SRMR	GFI	AGFI	$\Delta DF$	$\Delta\chi^2$	P
Factor Structure Equivalence (TF)	203.1 95	54	.000	3.763	.045 (.039-.052)	.077	.946	.910			
Factor Loading Equivalence	213.5 58	62	.000	3.444	.042 (.036-.049)	.078	.943	.917	8	10.363	.240
Factor Loading, Error Variances Equivalence	228.3 11	71	.000	3.216	.040 (.035-.046)	.077	.939	.923	9	14.753	.098

The  $\chi^2$  fit statistic for the factor loading equivalence model is 213.558 with 62 degrees of freedom. Subtracting the TF results from it produces the  $\Delta\chi^2$  value of 10.363 with 8 degrees of freedom, and with a *P* value of .240 (which is significantly distinguishable from 0 at  $P \geq .05$ ). This means that the added constraints do not significantly worsen the multi-group CFA model. Additionally, Table 5. 31 shows that, assuming the Factor Loading Equivalence model to be correct, the Error Variances Equivalence model did not significantly worsen model fit ( $P = .098$ ). Overall, the  $\chi^2/DF$ , RMSEA and AGFI values improved as the model became more constrained. The SRMR and the GFI stayed reasonably constant across all three models. It was clear from these figures that factor loading equivalence and error variances equivalence was present.

**5.3.2 PNS: Measurement Invariance for IRSP vs LTRS groups**

Table 5. 32 recaps the loose cross-validation fit statistics between the groups for the re-specified PNS measurement model.

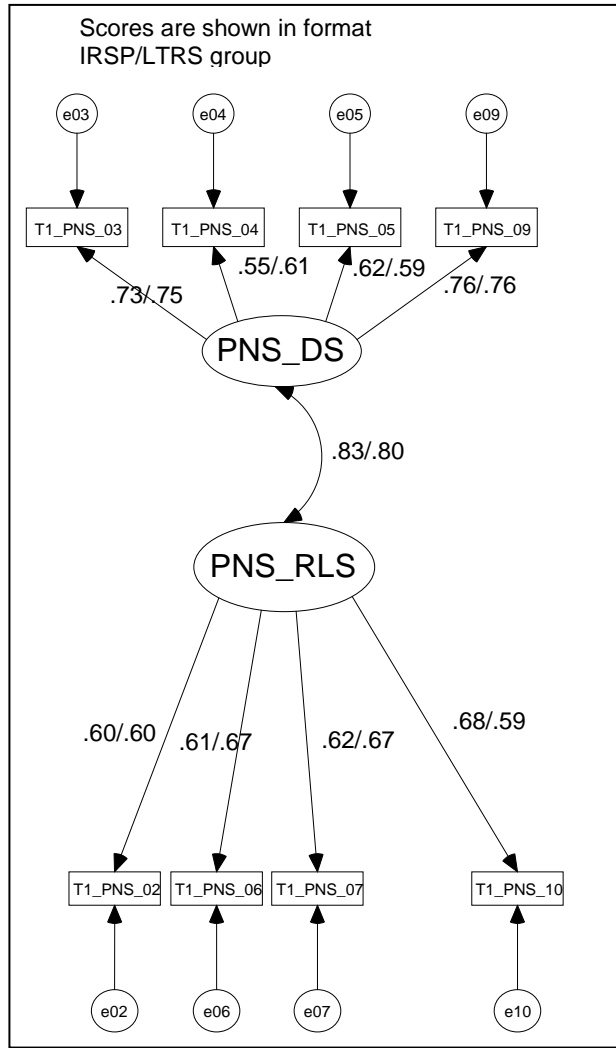
**Table 5. 32 Re-specified PNS model: Fit Statistics for IRSP and LTRS groups in T1**

Fit Index	IRSP	LTRS
N	777	583
$\chi^2$ (df)	42.687 (19)	61.178 (19)
p	.001	.000
$\chi^2/df$ ratio	2.247	3.220
GFI	.983	.978
AGFI	.968	.958
RMSEA (C.I.)	.040 (.024-.056)	.062 (.045-.079)
SRMR	.035	.045

### 5.3.2.1 *Factor Structure Equivalence*

Next, the same cross-validation steps were taken with the PNS measurement model, aside from the inclusion of an additional step testing inter-factor covariance equivalence, as this model possesses more than one latent factor.

The resulting  $\chi^2$  value was 103.872 with 38 degrees of freedom ( $p = .000$ ), and with a  $\chi^2/df$  of 2.733. The RMSEA for the multi-group model was 0.036 with a 90% confidence interval of .028 to .044, the SRMR was .035, the GFI was .98 and the AGFI was .962. These results support the multi-group PNS measurement model. Thus, factor structure equivalence was evidenced. Figure 5. 12 displays the resulting parameter estimates for each group. The estimated parameters generally support the model. However, several of the items had quite low estimates, although all were above the minimum .50 cut-off. Nonetheless, the model's overall fit was good.



**Figure 5. 12 PNS Re-specified model: Multi-group factor structure equivalence (TF model) parameter estimates.**

5.3.2.2 *Factor Loading Equivalence, Inter-factor Covariance and Error Variances*

*Equivalence*

Next, the CFA model was further constrained, step by step, only this time inter-factor covariance constraints were also applied. Table 5. 33 displays the fit statistics associated with all the equivalence models.

**Table 5. 33 Cross-Validation: Equivalence models for multi-group PNS in T1.**

Model	$\chi^2$	DF	P	$\chi^2/DF$	RMSEA (C.I.)	SRMR	GFI	AGFI	$\Delta DF$	$\Delta\chi^2$	P
Factor Structure Equivalence (TF)	103.8 72	38	.000	2.733	.036 (.028-.044)	.035	.980	.962			
Factor Loading Equivalence	112.3 31	44	.000	2.553	.034 (.026-.042)	.036	.979	.965	6	8.459	.206
Factor Loading and Inter-factor Covariance Equivalence	115.3 49	45	.000	2.563	.034 (.026-.042)	.042	.978	.965	1	3.018	.082
Factor Loading, Inter-factor Covariance, Error Variances Equivalence	134.4 21	53	.000	2.536	.034 (.027-.041)	.036	.974	.965	8	19.072	.014

Assuming Factor Structure Equivalence to be correct, the Factor Loading Equivalence model did not significantly worsen the fit ( $P = .206$ ). Assuming the Factor Loading Equivalence model to be correct, the addition of Inter-factor Covariance Equivalence also did not significantly worsen model fit ( $P = .082$ ). However, when Inter-factor Covariance Equivalence was assumed to be correct, applying Error Variances Equivalence did significantly worsen model fit as the P value fell below .05 ( $P = .014$ ). On the whole, all fit statistics stayed reasonably constant as the model became more constrained. It was clear from these figures that factor loading equivalence and inter-factor covariance equivalence was present, but error variances equivalence was not present.

**5.3.3 CoSI: Measurement Invariance for IRSP vs LTRS groups**

Table 5. 34 recaps the loose cross-validation fit statistics between the groups for the re-specified CoSI measurement model.

**Table 5. 34 Re-specified CoSI model: Fit Statistics for IRSP and LTRS groups in T1**

<b>Fit Index</b>	<b>IRSP</b>	<b>LTRS</b>
N	775	582
$\chi^2$ (df)	260.189 (74)	214.371 (74)
p	.000	.000
$\chi^2$ /df ratio	3.516	2.897
GFI	.909	.921
AGFI	.871	.888
RMSEA	.057	.057
(C.I.)	(.050-.065)	(.048-.066)
SRMR	.064	.069

### 5.3.3.1 Factor Structure Equivalence

Next, the same cross-validation steps were taken with the CoSI measurement model, again, with inter-factor constraints being introduced. The  $\chi^2$  value was 474.565 with 148 degrees of freedom ( $p = .000$ ), and with a  $\chi^2$ /df of 3.207. The RMSEA for the multi-group model was 0.040 with a 90% confidence interval of .036 to .044, the SRMR was .064, the GFI was .915 and the AGFI was .879. These results, although not as strong as with the other two (AO and PNS) measurement models, support the multi-group CoSI measurement model. Thus factor structure equivalence was evidenced. Figure 5. 13 displays the resulting parameter estimates for each group. The parameters estimated generally support the model. However, several of the items had quite low estimates (particularly items CoSI\_01, CoSI\_13 and CoSI\_15), although all were above the minimum .50 cut-off. Nonetheless, the model's overall fit was acceptable.



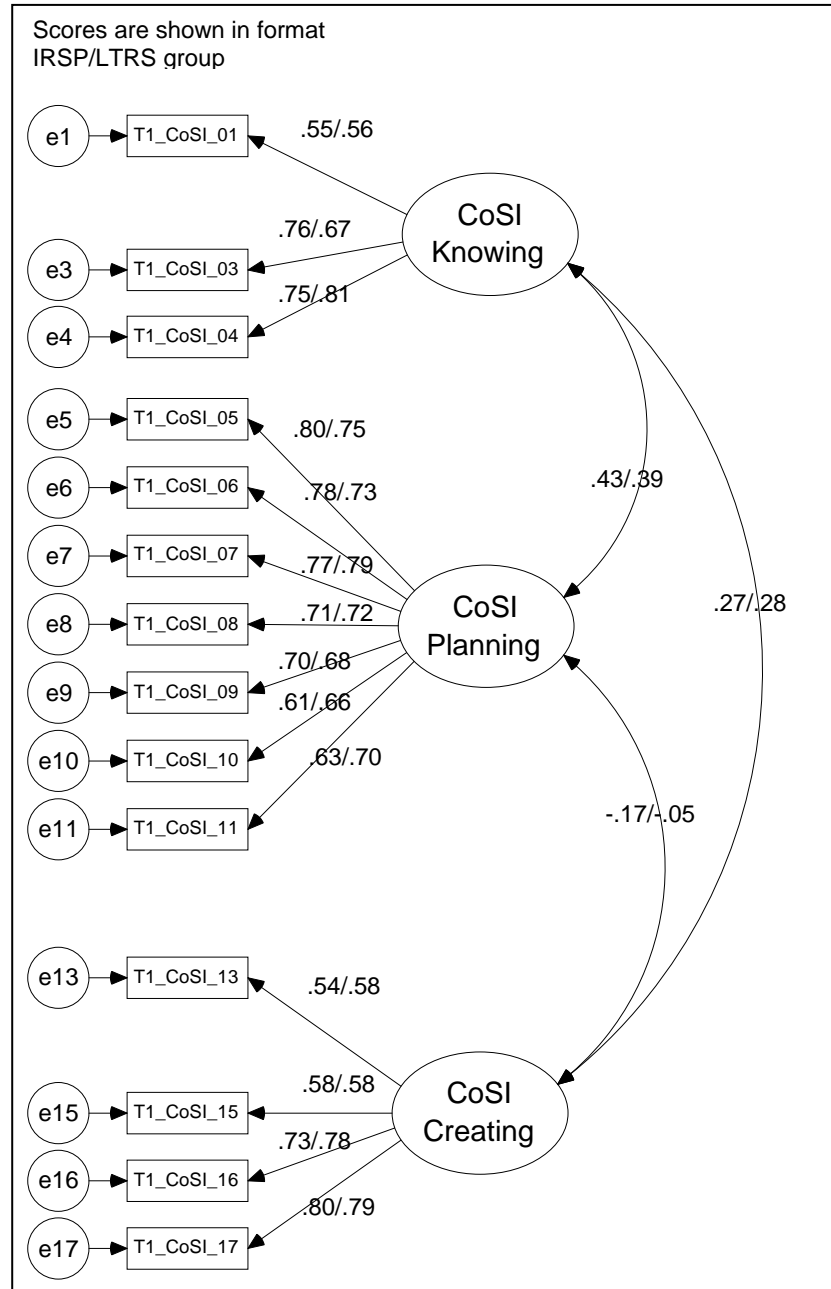


Figure 5. 13 CoSI Re-specified model: Multi-group factor structure equivalence (TF model) parameter estimates.

5.3.3.2 Factor Loading Equivalence, Inter-factor Covariance and Error Variances

Equivalence

Next, the CFA model was further constrained, step by step. Table 5. 35 displays the fit statistics associated with all the equivalence models.

**Table 5. 35 Cross-Validation: Equivalence models for multi-group CoSI in T1.**

Model	$\chi^2$	DF	P	$\chi^2/DF$	RMSEA (C.I.)	SRMR	GFI	AGFI	$\Delta DF$	$\Delta\chi^2$	P
Factor Structure Equivalence (TF)	474.5 65	148	.000	3.207	.040 (.036-.044)	.064	.915	.879			
Factor Loading Equivalence	482.6 37	159	.000	3.035	.039 (.035-.043)	.065	.913	.886	11	8.072	.707
Factor Loading and Inter-factor Covariance Equivalence	486.7 72	162	.000	3.005	.038 (.035-.042)	.065	.913	.887	3	4.135	.247
Factor Loading, Inter-factor Covariance, Error Variances Equivalence	529.6 22	176	.000	3.009	.039 (.035-.042)	.068	.905	.887	14	42.850	.000

Assuming Factor Structure Equivalence to be correct, the Factor Loading Equivalence model did not significantly worsen the fit ( $P = .707$ ). Assuming the Factor Loading Equivalence model to be correct, the addition of Inter-factor Covariance Equivalence also did not significantly worsen model fit ( $P = .247$ ). However, when Inter-factor Covariance Equivalence was assumed to be correct, applying Error Variances Equivalence constraints did significantly worsen model fit as the P value fell below .05 ( $P = .000$ ). On the whole, all fit statistics stayed reasonably constant as the model became more constrained. It was clear from these figures that factor loading equivalence and inter-factor covariance equivalence was present, but error variances equivalence was not present.

### 5.4 Stage 3: Testing for Validity and Reliability Across Time

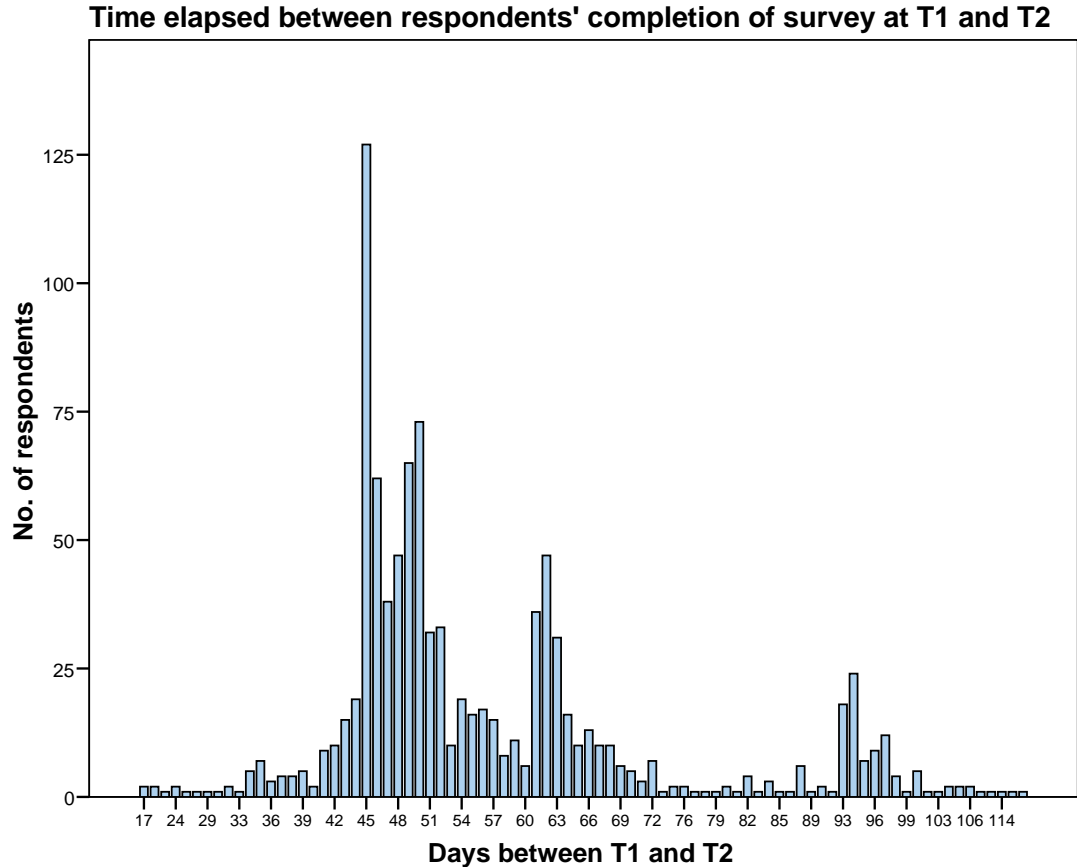
*Objective: To compare the test-retest reliability and validity of the IRSP group with the LTRS group.*

At this stage it was necessary to test the degree to which the data (for both groups) in time period 1 (T1) was replicated in time period 2 (T2).

### 5.4.1 Sample Considerations

Near the beginning of this chapter, Table 5. 1 presents the number of respondents in each test group in time period one, and the number that continued on to complete a repeat-measure in time period 2. Experimental mortality was noted for each test group along with the test-retest percentage of respondents (the percentage of total respondents that completed both parts). As can be seen from that table, the test-retest percentage was reasonably equivalent across test groups. In other words, experimental mortality was similar across groups, when taken as a percentage of the T1 sample.

Figure 5. 14 is a bar chart showing the number of days that elapsed between testing in time period 1 and time period 2, for all respondents. Peaks occur when the email invitations to complete part 2 were sent out, and where email reminders were sent out. The overall average number of days, between respondents completing the test in T1 and T2, was 56.6. This allowed enough time for an adequate wash-out period, without allowing too much time for any of the underlying psychometric constructs to change.



**Figure 5. 14 Time elapsed between T1 and T2 for all respondents.**

#### **5.4.2 Test-Retest Reliability**

As per Bagozzi (1994), any measurement can be thought of as an indicator of a theoretical concept, with *reliability* referring to the amount of agreement between independent attempts to measure the same theoretical concept. Highlighted, were the two key types of reliability; *internal consistency* and *test-retest reliability*. The former is present when two or more measures of the same theoretical concept, obtained at the same point in time, are in agreement, and was evidenced in Stage 1 of the analysis. The latter is present when measures of the same theoretical concept are repeated across time, with strong correlations between both sets of measures.

Ideally, in order to test for test-retest reliability, a SEM longitudinal (repeated-measures) model would have been implemented to test the data. Shown in Figure 5. 15 and Figure 5. 16, are the models that would have been specified in AMOS to test for this. However, the number of observed variables in these models have doubled with respect to models defined within Stage 2. One set of variables in the model represents time period 1, and the other time period 2. This has implications for fitting the model, given the necessary sample size for each test group needs to be larger than a necessary minimum when employing ADF estimation. Indeed, if one were to try running the models on the data, the repeat-measures AO model would run, however the CoSI model would not meet the minimum sample size requirements, with the following message being generated, “An error occurred while attempting to fit the model. A sample is too small for adf estimation. The sample has to exceed  $n*(n+1)/2$ , where n is the number of observed variables in the model”. In this case, the sample groups would need to be in excess of 406, whereas each of the four test groups have a sample size ranging from 202-297. Consequently, it was necessary to examine the test-retest reliability of both methods through a means other than this SEM technique.

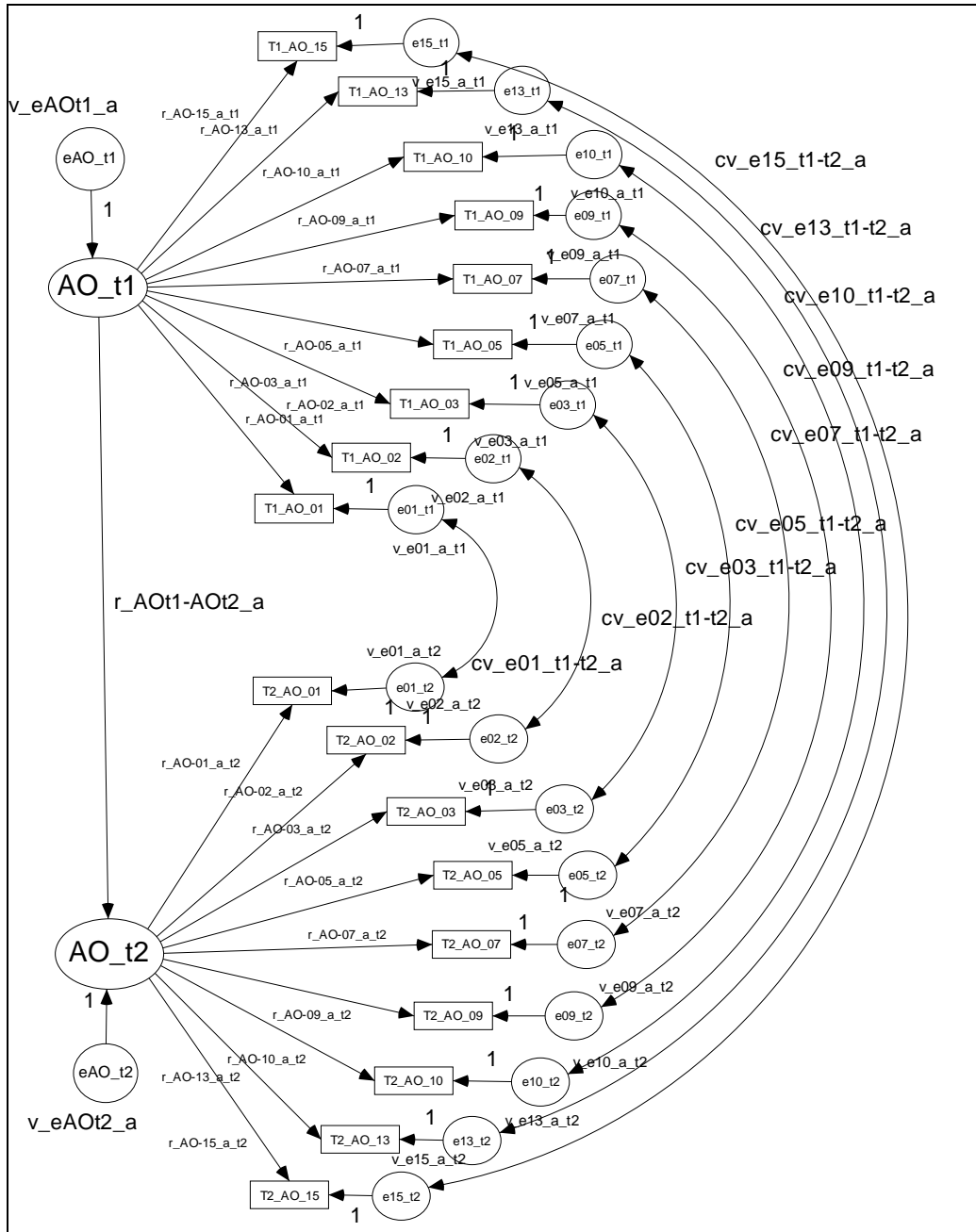


Figure 5.15 AO Repeat-measures structural equation model.

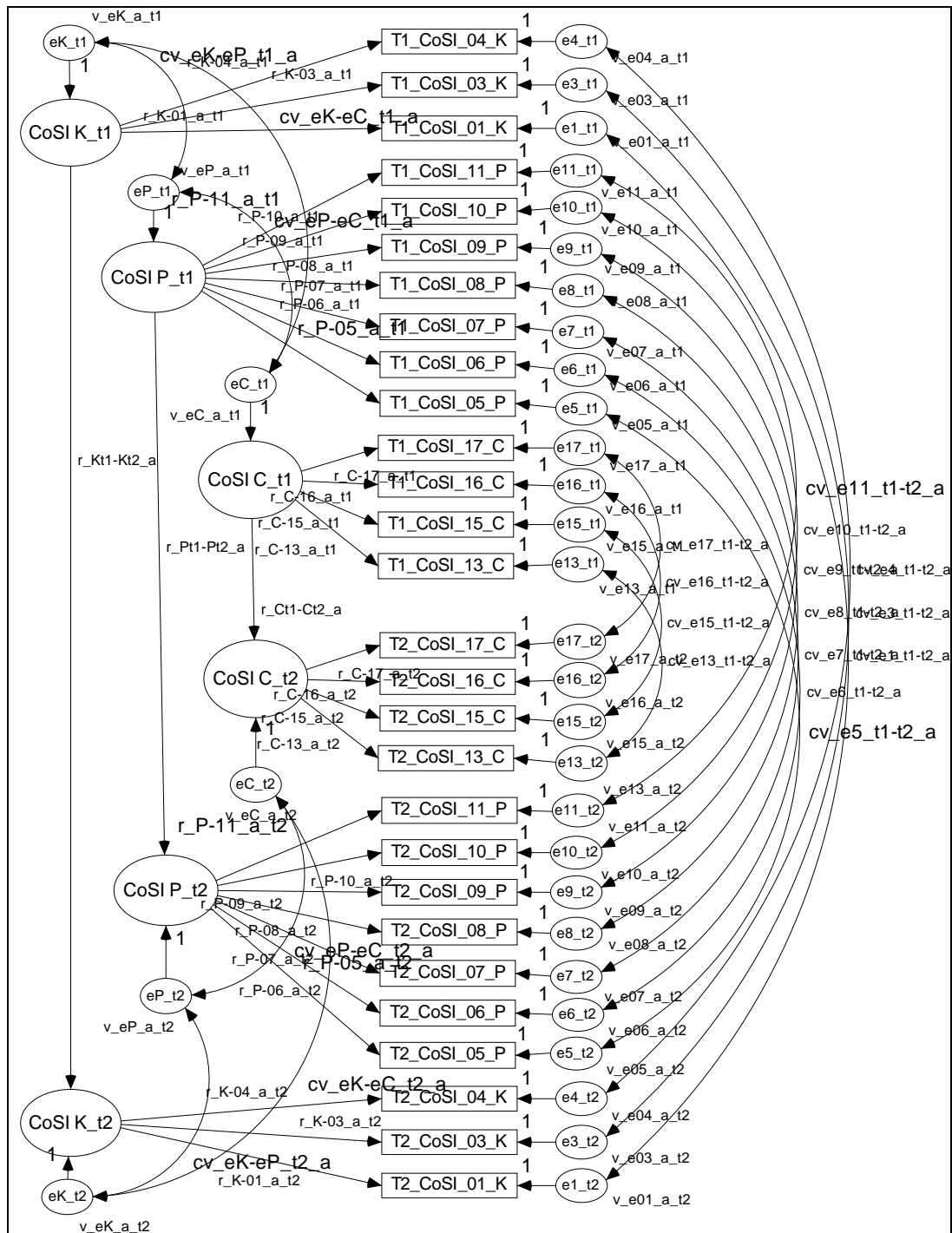


Figure 5. 16 CoSI Repeat-measures structural equation model.

Whilst these limitations meant that test-retest reliability could not be assessed using SEM, it could, however, be determined through correlation tests between psychometric factor scores from T1 to T2. Ideally, Pearson correlations would have been employed. However, as has already been established, the distribution of the data on the

psychometric variables was extremely non-normal. As such, Spearman correlations were conducted on factor scores for AO, PNS, CoSI and BFI, between the two time periods.

Before this could be done, the factor scores needed to be computed from the data in SPSS. Stage 1 and 2 established the measurement models and their associated levels of equivalence. Any factor scores, used henceforth, were based on the item weightings from the measurement models at the highest level of equivalence achieved. For the BFI factors, summated scores were used. Subsequent to the computation of these factor scores, a test of outliers was conducted on these scores as per the SPSS method outlined in section 5.2.1 of Stage 1. Each factor was examined for outliers by that method, and a total of 42 outliers were removed. Table 5. 36 details the new sample sizes for the groups.

**Table 5. 36 Sample sizes for groups T1 and T2, after the removal of all outliers by factor scores.**

<b>Test Group</b>	<b>Treatments</b>	<b>No. respondents that completed T1</b>	<b>No. respondents that completed T1 &amp; T2</b>
TG <sub>1</sub>	IRSP-IRSP	364	263
TG <sub>2</sub>	IRSP-LTRS	377	284
TG <sub>3</sub>	LTRS-IRSP	284	194
TG <sub>4</sub>	LTRS-LTRS	281	203
	<b>Totals</b>	<b>1306</b>	<b>944</b>

Spearman correlations were conducted on TG1 (the IRSP-IRSP group) and TG4 (the LTRS-LTRS group). This permitted a comparison of methods on the test-retest reliability scores across the factors. Although this provided a less powerful measure of test-retest reliability than that which would have been given through SEM, it is still a



good indicator. Table 5. 37 shows the test-retest reliability values (the correlations) for each of the factors, by method.

**Table 5. 37 Spearman correlations between T1 and T2 psychometric factor scores: Test-retest reliability by method (TG1 vs TG4).**

	AO	PNS_ DS	PNS_ RLS	CoSI _C	CoSI _K	CoSI _P	Ext	Agr	Con	Neu	Ope
<b>IRSP</b> N=263	<b>.772</b>	.711	.721	<b>.691</b>	.584	<b>.712</b>	<b>.842</b>	<b>.714</b>	.683	.737	.709
<b>LTRS</b> N=232	.716	<b>.776</b>	<b>.785</b>	.657	<b>.655</b>	.670	.836	.624	<b>.749</b>	<b>.789</b>	<b>.731</b>

All correlations were significant at the .01 level, two-tailed

The IRSP group achieved higher test-retest reliability for the AO scale than the LTRS group. However, the situation was reversed for the PNS scale, with the LTRS group producing higher values for reliability on both factors. On the whole, reliability scores were lower on all three of the CoSI factors, for both groups. However, two out of the three factors achieved higher reliability levels in the IRSP group. Results are mixed again with the five factors of the BFI, with three out of the five factors scoring higher reliability values with the LTRS group. These results do not appear to favour one method over the other. In fact, both groups performed similarly well.

Shifting the focus solely on the IRSP group, a further test of reliability was conducted on those who defined exactly the same IRS in both time periods (i.e. same number of intervals, and degree of balance), versus those who defined different IRSs in each time period. The results are outlined in Table 5. 38.

**Table 5. 38 Spearman correlations between T1 and T2 psychometric factor scores: Test-retest reliability for TG1 (IRSP-IRSP), by IRS change.**

<b>IRSP T1-T2</b>	<b>AO</b>	<b>PNS_ DS</b>	<b>PNS_ RLS</b>	<b>CoSI _C</b>	<b>CoSI _K</b>	<b>CoSI _P</b>	<b>Ext</b>	<b>Agr</b>	<b>Con</b>	<b>Neu</b>	<b>Ope</b>
<b>IRS same N=105</b>	.747	.667	.689	.657	<b>.641</b>	.661	<b>.869</b>	<b>.731</b>	<b>.718</b>	.664	.680
<b>IRS change N=158</b>	<b>.801</b>	<b>.742</b>	<b>.738</b>	<b>.698</b>	.546	<b>.743</b>	.817	.705	.649	<b>.777</b>	<b>.724</b>

All correlations were significant at the .01 level, two-tailed

Interestingly, seven out of the eleven factor scores produced higher correlations in the group that defined *different* IRSs the second time round. This means that despite that fact that they chose to define a different IRS, on the whole, the test-retest reliability of their scores were equally as good as those who kept their IRSs the same, as well as those from the LTRS group. The results are not significantly in favour of one group over the other.

**5.4.3 Multitrait-Multimethod (MTMM) Matrix Extending Validity Testing**

Whilst construct validity (convergent and discriminant validity) for the psychometric constructs was examined in Stage 1 for each individual measurement model, they could not all be examined in a single measurement model (i.e. with CoSI, PNS and AO factors all together) given the number of variables would have demanded a much larger sample (under ADF estimation). Moreover the BFI two-item factors could not be examined properly in SEM, given the measurement model would have been under-identified. For this reason, Campbell and Fiske’s multitrait-multimethod (MTMM) matrix approach was used as a further examination of validity. This meant that the BFI factors could be examined, so too could the convergent and discriminant validity between *all* the factors, *across* the models. To examine construct validity, Campbell and Fiske (1959) proposed that *convergent* and *discriminant validity* be examined by analysing the correlations

produced by measuring two or more concepts (which they termed “traits”) by two or more methods. The MTMM matrix approach presents the correlation matrix between measures of traits by methods. Campbell and Fiske’s (1959) paper and Bagozzi’s (1994) book provide a detailed outline of Campbell and Fiske’s MTMM matrix and its application. Their original 3x3 method matrix has been illustrated here as a 2x2 method matrix (given its appropriateness), and is shown in Table 5. 39.

**Table 5. 39 Campbell and Fiske’s MTMM Matrix.**

Traits		Method 1			Method 2		
		A <sub>1</sub>	B <sub>1</sub>	C <sub>1</sub>	A <sub>2</sub>	B <sub>2</sub>	C <sub>2</sub>
<b>Method 1</b>	A <sub>1</sub>	<i>(r<sub>A<sub>1</sub>A<sub>1</sub></sub>)</i>					
	B <sub>1</sub>	<i>r<sub>B<sub>1</sub>A<sub>1</sub></sub></i>	<i>(r<sub>B<sub>1</sub>B<sub>1</sub></sub>)</i>				
	C <sub>1</sub>	<i>r<sub>C<sub>1</sub>A<sub>1</sub></sub></i>	<i>r<sub>C<sub>1</sub>B<sub>1</sub></sub></i>	<i>(r<sub>C<sub>1</sub>C<sub>1</sub></sub>)</i>			
<b>Method 2</b>	A <sub>2</sub>	<i>r<sub>A<sub>2</sub>A<sub>1</sub></sub></i>	<i>r<sub>A<sub>2</sub>B<sub>1</sub></sub></i>	<i>r<sub>A<sub>2</sub>C<sub>1</sub></sub></i>	<i>(r<sub>A<sub>2</sub>A<sub>2</sub></sub>)</i>		
	B <sub>2</sub>	<i>r<sub>B<sub>2</sub>A<sub>1</sub></sub></i>	<i>r<sub>B<sub>2</sub>B<sub>1</sub></sub></i>	<i>r<sub>B<sub>2</sub>C<sub>1</sub></sub></i>	<i>r<sub>B<sub>2</sub>A<sub>2</sub></sub></i>	<i>(r<sub>B<sub>2</sub>B<sub>2</sub></sub>)</i>	
	C <sub>2</sub>	<i>r<sub>C<sub>2</sub>A<sub>1</sub></sub></i>	<i>r<sub>C<sub>2</sub>B<sub>1</sub></sub></i>	<i>r<sub>C<sub>2</sub>C<sub>1</sub></sub></i>	<i>r<sub>C<sub>2</sub>A<sub>2</sub></sub></i>	<i>r<sub>C<sub>2</sub>B<sub>2</sub></sub></i>	<i>(r<sub>C<sub>2</sub>C<sub>2</sub></sub>)</i>

Note: The validity diagonals are the set of italicised values. The reliability diagonals are the set of values in parentheses. Each heterotrait-monomethod triangle is enclosed by a solid line. Each heterotrait-heteromethod triangle is enclosed by a broken line.

The key point is that the matrix is used to assess convergent and discriminant validity in four steps:

The first step provides evidence of convergent validity:

1. Entries in the validity diagonal should be significantly different from 0 and sufficiently large to encourage further examination of validity.

The next three steps provide evidence of discriminant validity:

2. A validity diagonal value should be higher than that of the values lying in its column and row in the heterotrait-heteromethod triangles. That is, a validity value for a variable should be higher than the correlations obtained between that variable and any other variable having neither trait nor method in common.

3. A variable should correlate higher with an independent effort to measure the same trait than with measures designed to get at different traits which happen to employ the same method. For a given variable, this involves comparing its correlation values in the validity diagonals with those in the heterotrait-monomethod triangles.
4. The same pattern of trait interrelationship should ideally be shown in all of the heterotrait triangles of both the monomethod and heteromethod blocks.

It is necessary to point out that the MTMM approach assumes that respondents are measured using both methods in the same time period. In the current study, this would have been inappropriate given; (a) the length of the survey would have doubled (thus reducing participation rates), (b) bias would have been introduced through fatigue effects and the lack of a wash-out period between test methods, and (c) this would not have provided a suitable test design to examine test-retest reliability. The key adaptations are illustrated by Table 5. 40 and Table 5. 41.

**Table 5. 40 Original MTMM Matrix approach.**

	Method 1	Method 2
Method 1	Correlations between the traits, measured using Method 1.	
Method 2	Correlations between the traits, measured using Method 2 against those measured using Method 1.	Correlations between the traits, measured using Method 2.

**Table 5. 41 Adapted MTMM Matrix approach.**

	IRSP	LTRS
IRSP	Time period 1 Correlations between the traits, measured using IRSP. Sample used: Test Group 1	
LTRS	Time period 1 – Time period 2 Correlations between the traits, measured using LTRS in T1, against those measured using IRSP in T2. Sample used: Test Group 3	Time period 1 Correlations between the traits, measured using LTRS. Sample used: Test Group 4

Note: Test Group 1 (IRSP-IRSP).  
 Test Group 3 (LTRS-IRSP).  
 Test Group 4 (LTRS-LTRS).

The IRSP-IRSP and the LTRS-LTRS quadrants could be tested in the traditional MTMM manner, whereby correlations between the traits are calculated for the same method (in a single time period). With respect to the heteromethod quadrant (LTRS-IRSP), given respondents were likely to have encountered Likert-type rating-scales in the past but will be unfamiliar with the IRSP method of measurement, it was thought prudent to examine T1-T2 correlations from measurements captured using LTRS in the first instance and the IRSP in the second. This ensured that any possible ‘novelty’ effect (from using the new IRSP method first) on the repeat-measure (LTRS), did not enter into the analysis. For these reasons TG3 (LTRS-IRSP) was chosen over TG2 (IRSP-LTRS) for examination in the heteromethod quadrant. Table 5. 42 shows the resultant (Spearman) correlations for the MTMM matrix for the IRSP and LTRS data.

On examination of the Cronbach alphas for the AO, PNS, and CoSI factors, (Table 5.42), they indicate that reliability (internal consistency) is present<sup>8</sup>, which supports the SEM results from Stage 1. On closer examination of the IRSP's performance (top-left quadrant), reliability is very high for the AO construct (.930), and acceptable for the PNS (.731, .705) and CoSI constructs (.725, .846, .747). However, the Cronbach alphas are somewhat weak for the five BFI constructs (.645, .403, .498, .551, .495). These five BFI constructs may have been more prone to internal consistency problems, given each was measured with only two items. In fact, when examining the LTRS' coefficients (bottom-right quadrant), the Cronbach alpha scores have a very similar pattern to that in the IRSP quadrant. As for the other LTRS coefficients, the AO construct had very high reliability (.905), with all but one of the PNS and CoSI constructs achieving acceptable levels (ranging from .730-.831). This one exception was the CoSI\_K construct, which had a reliability coefficient of .648. Under the IRSP measure (top-left quadrant), this same construct achieved a more acceptable reliability level of .725. Interestingly, the five BFI constructs also achieved low levels of reliability under the LTRS measure, ranging from .230-.530. The BFI Agreeableness construct had a coefficient alpha of .230, which was extremely low, considering the LTRS used was that pre-validated in the literature for the scale (as with all the constructs). The IRSP measure achieved a higher reliability coefficient for this construct (.403), but this too was still inadequate.

---

<sup>8</sup> The coefficients achieved the desired minimum .7 level (Bagozzi, 1994).

**Table 5. 42 MTMM Matrix: IRSP and LTRS.**

Methods		IRSP											LTRS										
Traits		AO	PNS DS	PNS RLS	CoSI K	CoSI P	CoSI C	BFI Ex.	BFI Ag.	BFI Co.	BFI Ne.	BFI Op.	AO	PNS DS	PNS RLS	CoSI K	CoSI P	CoSI C	BFI Ex.	BFI Ag.	BFI Co.	BFI Ne.	BFI Op.
IRSP	AO	<b><u>0.930</u></b>																					
	PNS DS	-0.004	<b><u>0.731</u></b>																				
	PNS RLS	-0.010	0.558	<b><u>0.705</u></b>																			
	CoSI K	-0.100	0.070	0.001	<b><u>0.725</u></b>																		
	CoSI P	-0.004	0.462	0.333	0.293	<b><u>0.846</u></b>																	
	CoSI C	0.052	-0.307	-0.408	0.206	-0.088	<b><u>0.747</u></b>																
	BFI Ex.	0.056	-0.195	-0.250	-0.038	-0.140	0.177	<b><u>0.645</u></b>															
	BFI Ag.	0.087	-0.018	-0.115	-0.020	0.036	0.048	0.066	<b><u>0.403</u></b>														
	BFI Co.	-0.122	0.165	-0.014	0.245	0.331	0.033	0.056	0.018	<b><u>0.498</u></b>													
	BFI Ne.	0.145	0.152	0.364	0.021	0.189	-0.185	-0.255	-0.169	-0.146	<b><u>0.551</u></b>												
BFI Op.	0.106	-0.216	-0.198	0.100	-0.125	0.209	0.103	-0.056	0.050	0.071	<b><u>0.495</u></b>												
LTRS	AO	0.746	0.087	0.106	-0.071	0.043	-0.102	0.058	0.048	-0.040	0.076	0.100	<b><u>0.905</u></b>										
	PNS DS	0.066	0.706	0.464	0.076	0.379	-0.303	-0.165	0.016	0.154	0.199	-0.122	0.003	<b><u>0.736</u></b>									
	PNS RLS	0.152	0.535	0.744	0.000	0.267	-0.287	-0.362	-0.134	-0.107	0.359	-0.048	0.023	0.530	<b><u>0.730</u></b>								
	CoSI K	-0.011	0.142	-0.005	0.500	0.217	0.140	-0.077	-0.032	0.168	0.020	0.134	-0.151	0.036	-0.112	<b><u>0.648</u></b>							
	CoSI P	-0.039	0.593	0.390	0.134	0.580	-0.119	-0.195	0.100	0.220	0.267	-0.044	0.025	0.480	0.353	0.186	<b><u>0.831</u></b>						
	CoSI C	-0.147	-0.354	-0.399	0.157	-0.193	0.625	0.272	0.171	0.049	-0.221	0.202	0.038	-0.241	-0.310	0.234	-0.089	<b><u>0.788</u></b>					
	BFI Ex.	0.098	-0.237	-0.364	0.028	-0.073	0.228	0.852	0.131	0.101	-0.383	0.045	0.110	-0.218	-0.370	-0.024	-0.008	0.218	<b><u>0.675</u></b>				
	BFI Ag.	-0.021	-0.086	-0.167	-0.080	0.016	0.100	0.088	0.716	0.130	-0.171	-0.073	0.053	-0.006	-0.117	-0.058	-0.067	-0.024	0.123	<b><u>0.230</u></b>			
	BFI Co.	0.049	0.004	-0.055	0.184	0.214	0.109	0.185	0.161	0.751	-0.104	0.129	-0.027	0.194	0.075	0.214	0.260	-0.016	0.118	0.098	<b><u>0.508</u></b>		
	BFI Ne.	0.163	0.317	0.420	-0.094	0.153	-0.249	-0.369	-0.130	-0.116	0.838	0.107	0.103	0.121	0.416	-0.086	-0.006	-0.194	-0.278	-0.246	-0.183	<b><u>0.530</u></b>	
BFI Op.	0.133	-0.085	0.024	0.166	-0.004	0.269	0.026	0.015	0.079	0.097	0.680	0.111	-0.133	-0.116	0.095	-0.018	0.276	0.091	-0.021	-0.023	0.063	<b><u>0.514</u></b>	

The figures emboldened and underlined show Cronbach's alpha<sup>9</sup> for the construct. Shaded figures are the correlations that were significant at the 0.01 level (2-tailed).

<sup>9</sup> A general rule of thumb for a reliability estimate is that .7 or higher suggests good reliability, and between .6 and .7 are considered acceptable (Hair et al., 2006).

An inspection of the correlations in Table 5. 42 shows first that convergent validity is achieved in that all four monotrait-heteromethod correlations from the validity diagonal (i.e. .746, .706, .744...etc) are large and significantly different from zero ( $p < 0.01$ ).

The first discriminant validity criterion was reasonably satisfied as each monotrait-heteromethod correlation was greater than each correlation lying in its row and column of the heterotrait-heteromethod triangles, except for one. There was a single monotrait-heteromethod correlation, CoSI P – CoSI P (.580), that was not larger than *all* the heterotrait-heteromethod correlations in its row and column, as there was a single correlation that was larger, CoSI P – PNS DS (.593). Aside from this one, all the others met the first of the three discriminant validity criteria. This single violation, given the proportion of comparisons that passed, can be considered anomalous (Bagozzi, 1994), with the matrix still meeting the first criterion for discriminant validity due to multiple comparisons. Specifically, the criterion involved a total of 220 comparisons (219 of which passed) for this eleven-trait, two-method matrix, and all the differences in correlations were statistically significant ( $p < .05$ ). As such, the first criterion for discriminant validity was satisfied.

Application of the second discriminant validity criterion showed that all the validity diagonal correlations were greater than their corresponding entries in the heterotrait-monomethod triangles. For all 220 comparisons, differences in correlations were statistically significant and in the proper direction.

Finally, the patterns of correlations in all heterotrait triangles in both monomethod and heteromethod instances were compared by use of Kendall's Coefficient of



Concordance. The hypothesis for the pattern of correlations was found to be discordant:  $\chi^2 (54) = 168.221$ ,  $p < 0.001$ , with Kendall's  $W = .779$ , which indicates a strong agreement between the patterns of correlations of the four triangles. A key point is that this means that the pattern of inter-factor correlations yielded from data collected using the LTRS method, was statistically similar to the pattern from the data collected using the IRSP method.

To summarise, the data in Table 5. 42 show that the measures of the AO, CoSI and PNS traits achieved both convergent and discriminant validity for both methods. Given the weak reliability (internal consistency) coefficients for the five BFI constructs, they could not be shown to possess validity for either IRSP or LTRS. However, what was most important to draw from this was that the results obtained using the IRSP method were comparable to those obtained using the LTRS.

## **5.5 Stage 4: Individualised Rating-Scales (IRSs) and Individual**

### **Characteristics**

*Objective: To examine whether there are any relationships between respondents' IRSs and their individual characteristics.*

#### **5.5.1 Sample**

In order to address this objective, the data from time period 1 (T1) was used for much of the analysis. Where appropriate, comparisons were made with time period 2 data. The number of sample units obtained in T1 was more than sufficient to provide adequate power to the necessary tests, and the exclusion of data from T2 from some of the tests

would mean that any potential test-retest carry-over contamination effects, whilst minimised through a suitable wash-out period, will be avoided.

In addition, given the focus here was to examine the data for any relationships between specific variables, it was considered desirable to increase internal validity as much as possible and minimise unwanted variation in the dependent variables caused by differences in extraneous independent variables such as ‘first language’. It was therefore worthwhile examining the data to see whether the sample could be made even more homogenous, whilst maintaining an adequate sample size for the planned analyses. As shown in Table 5. 43, out of a total of 741 respondents that completed the IRSP survey in T1, 554 were British, and 538 of those spoke English as their first language. This group of 538 respondents was used as the basis for all analyses in this stage. Therefore, in this section (5.5), whenever references are made to respondents that completed the IRSP in T1, it is the 538 British and English-speaking respondents that are being referred to. Streamlining the sample by Language and by Nationality in this way permitted a greater degree of homogeneity, whilst still maintaining an adequate sample size.

**Table 5. 43 IRSP Respondents in T1: By national identity and first language.**

		National Identity						Total
		British	European	North American	Dual	Other	Unknown	
First Language	English	<b>538</b>	2	63	27	28	8	666
	Other	16	24	1	6	21	7	75
Total		554	26	64	33	49	15	741

### 5.5.2 IRS Lengths Chosen

First and foremost, the rating-scale lengths chosen by respondents were examined. The IRSP process involved having respondents define their IRS, then use it to rate sixteen uncorrelated items, followed by receiving an option to revise their IRS. This means that respondents who chose to revise their IRSs before proceeding onto the main survey, will have had two IRSs registered by the survey. This permitted an examination of whether or not respondents required the trial run-through of their IRS, and whether they underwent a ‘learning’ process. This means that in any one time period (in this case T1), the term IRSL\_1 is used when referring to the length of respondents’ first IRS, and the term IRSL\_2 is used when referring to their second. The term IRSL\_Used refers to the length of the IRS that the respondent used when completing the main survey; which would be IRSL\_1 for those who opted out of changing their IRS, and IRSL\_2 for those who chose to modify it.

Figure 5. 17 illustrates the spread of the IRSL\_Used for respondents (the rating-scale lengths ultimately chosen) in a histogram. The mean IRSL\_Used was 8.99 with a standard deviation of 3.33. It is quite clear from the graph that the mode rating-scale length was one with seven categories. The smallest was a three-category IRS, the smallest allowable, in that it consisted only of neutral and two endpoints. The largest value was a 23-category IRS (the maximum allowable). It is unsurprising, given the qualitative insights, that most respondents opted for a balanced bipolar rating-scale (the same number of categories on both sides of the continuum), clearly seen on the graph where the odd numbers score higher than the even ones. Of these respondents, 83%

defined balanced IRSs, meaning that approximately one in five people prefer to have an imbalanced rating-scale.

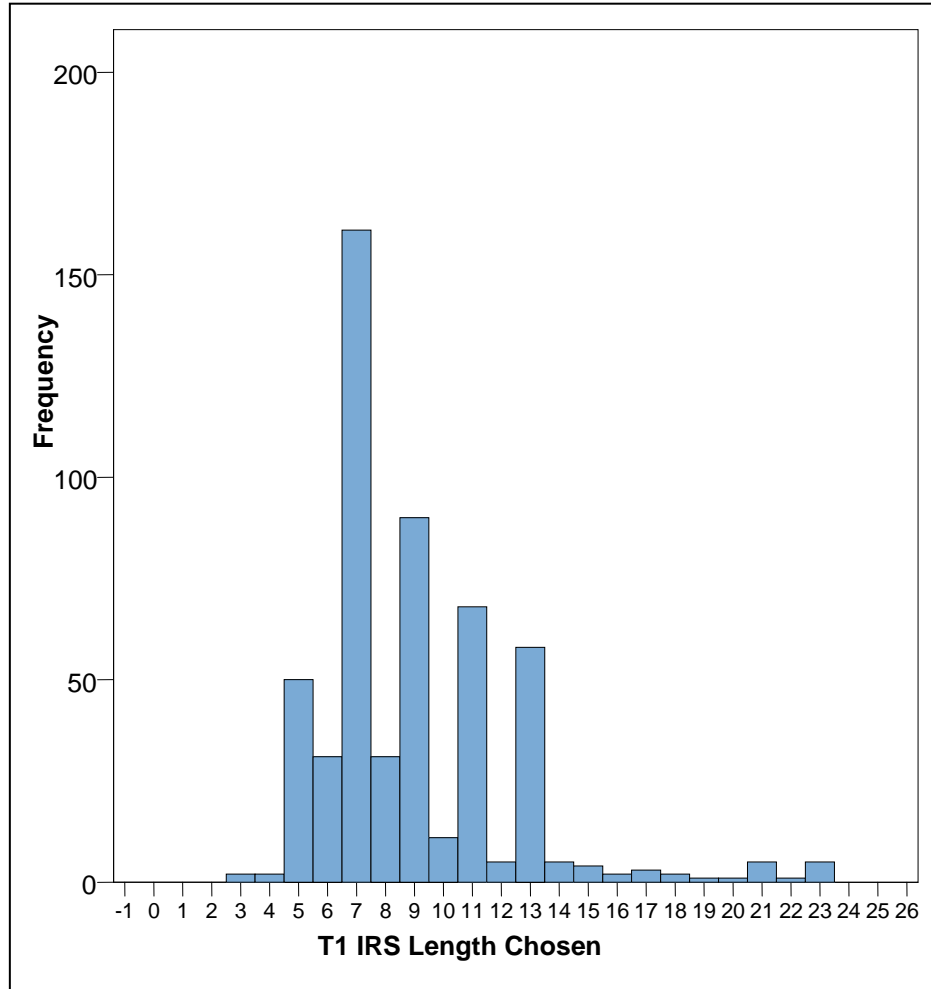


Figure 5. 17 T1 IRS Lengths Used

Of the 538 respondents, 116 opted to modify their IRS length after having practiced using it on the sixteen uncorrelated items. That is approximately 22% who will have experienced some form of within-survey learning<sup>10</sup> and decided to modify their IRS before proceeding further. Thus, it would seem that the option, which provides respondents with an opportunity to modify their IRS, is a necessary one given approximately one in five took it.

Of those who completed the IRSP in T1, 192 of those invited, returned to complete the IRSP in T2. Of the 116 respondents who modified their IRS length in T1, 46 completed the IRSP survey in T2. A paired sample t-test was conducted<sup>11</sup> on those 46 respondents to see whether their IRSL\_2 (effectively their IRSL\_Used) from T1 was significantly different to their IRSL\_1 in T2. In other words, this test was done to see whether the modified version of their IRS in T1 was replicated the first time they defined their IRS some time later in T2. The results showed that respondents' IRSL\_2 in T1 (mean = 8.80; SD = 2.99) were not significantly different from their IRSL\_1 in T2 (mean = 8.80; SD = 2.43);  $t(df = 45) = .000, p = 1.00, CI = -1.068 \text{ to } 1.068$ . It would therefore seem that the within-survey learning that these respondents experienced in T1 was taken through to T2. This stability, between IRSL\_2 in T1 and IRSL\_1 in T2, would also suggest that the length of their modified IRS in T1 (IRSL\_2) was personally appropriate, in that these respondents chose to define the very same IRS length the next time they completed the survey. In fact, for those who modified their IRS in T1, a paired sample t-test found no significant difference between the IRSL\_Used in T1 (mean = 8.80; SD = 2.99) and the final IRSL\_Used in T2 (mean = 8.37; SD = 2.27);  $t$

---

<sup>10</sup> They practiced using their IRS before opting to modify it, and therefore experienced 'learning' *during* the survey in T1.

<sup>11</sup> Note, although the data has been found to be non-normally distributed, the validity of the paired-sample t test is not compromised because the normality assumption can be ignored where  $n > 30$  (Gravetter and Wallnau, 2004).

(df = 45) = .889, p = .379, CI = -.550 to 1.420. This addressed the possibility that respondents could have modified their IRSs again in T2. It would seem that once respondents adopted their modified IRS in T1, it appeared to be stable even after a wash-out period.

As for those who did not choose to modify their IRS length in T1, there was no significant difference between their IRSL\_Used (which would be their IRSL\_1) in T1 (mean = 8.99; SD = 3.47), and their IRSL\_Used in T2 (mean = 8.55; SD = 3.33); t (df = 145) = 1.863, p = .064, CI = -.026 to .889. This would suggest that even for those who felt they had defined an appropriate IRS length first time (in T1), it was kept again next time they used the IRSP method.

Table 5. 44 is a crosstabulation showing the proportion of those who chose to modify/not modify their IRS length in T1, against their choice in T2.

**Table 5. 44 IRS length: Choice to modify, by respondents from T1 to T2.**

		T2_IRS Length		Total
		no modification	modified length	
T1_IRS Length	no modification	138	8	146
	modified length	37	9	46
Total		175	17	<b>192</b>

Whilst there is a significant reduction from those who chose to modify their IRS length in T1, to those who chose to modify in T2, some chose to modify in both time periods (9 respondents out of 46 to be precise). That is approximately 20%. The other 80% appeared to have ‘learned’ from the first modification experience of T1, and chose not to modify their IRS length in T2. Interestingly, approximately 5% who did not modify their IRS length in T1 chose to do so in T2.

### 5.5.2.1 *Individual Characteristics and IRS Length*

The key individual characteristics captured by the survey consisted of demographic characteristics (gender, postgraduate/undergraduate status, degree subject, first language, national identity), and personal traits (Affective Orientation; Personal Need for Structure: Desire for Structure and Response to Lack of Structure; Cognitive Styles: Knowing, Planning and Creating; the Big Five Personality types: Extraversion, Conscientiousness, Neuroticism, Openness and Agreeableness). One of the objectives of this study was to examine whether any relationships exist between individual characteristics and choice of IRS length. Given that first language and national identity are English and British respectively for this sample of 538 respondents who completed the IRSP in T1, these two demographic characteristics are ignored. Even if the other categories of these demographic characteristics had been included, they would have contained too small a number of respondents to be representative of the groups concerned. Of the demographic characteristics, this leaves gender, postgraduate/undergraduate status, and degree subject, to examine.

Ideally, both the demographic characteristics and the individual trait measures would have been included in a multiple regression model whereby all personal characteristics would have been the independent variables (with the demographic variables assigned dummy categories), and the dependent variable would have been IRS length. However, as has already been established, the psychographic variables in this data set are all severely non-normally distributed. This would render the results from any regression analysis untrustworthy given that normality assumptions would be violated. Given this is an exploratory methodological study, it was not necessary to be able to determine

conclusively what individual traits predict IRS length. However, wherever possible within the constraints set by the data, any discoveries as to possible associations between individual characteristics and IRS length were considered of value. As such, the parametric variables (i.e. individual traits) were examined using Spearman's correlation, given it is robust to skewed data. The nonparametric variables (i.e. the demographic characteristics) were examined for differences using the Mann-Whitney U-test and the Kruskal-Wallis test for multiple categories, which are also robust to violations of normality. Mean scores were examined where significant differences were found to exist between groups on the demographic variables.

First, males and females (the independent variables) were examined for differences in their choice of IRS length (the dependent variable). The IRS lengths were rank-ordered and the Mann-Whitney U-test was used to compare the ranks for the males ( $n=164$ ) versus the females ( $n=374$ ). The results indicated no significant difference between the genders,  $U = 29129$ ,  $p=.345$ , with the sum of the ranks equal to 45737 for males and 99254 for females.

Next the Mann-Whitney U-test was used to compare the ranks for the undergraduates ( $n=421$ ) versus the postgraduates ( $n=117$ ). The results indicated no significant difference between the two groups,  $U = 24151$ ,  $p=.744$ , with the sum of the ranks equal to 113937 for undergraduates and 31054 for postgraduates.

Finally the Kruskal-Wallis test was used to evaluate differences among the ten categories of degree area, shown in Table 5. 45. The outcome of the test indicated no significant differences among the groups,  $H = 9.791$  ( $df=9$ ,  $N=535$ ),  $p=.368$ . This result



suggests that there are no differences between how respondents studying different disciplines chose to define their IRS length.

**Table 5. 45 T1 IRSP: Kruskal-Wallis Rankings for IRS Length by Degree Area.**

Degree_Area	N	Mean Rank
Arts	10	252.40
Business	38	249.29
Engineering	18	252.78
History	69	308.32
IT	16	237.56
Languages	111	260.72
Law	50	241.06
Life Sciences	116	261.28
Physical Sciences	28	278.00
Social Sciences	79	286.99
<b>Total</b>	<b>535*</b>	

\* 3 respondents' degree areas were unknown which is why n=535 and not 538.

All these results suggest that there is no observable relationship between respondents' demographic characteristics and their choice of IRS length.

Next, individual traits were examined for any associations with IRS length. Spearman correlations were obtained for IRSL\_Used in T1 and each of the individual trait scores from T1: extraversion, conscientiousness, agreeableness, neuroticism, openness, desire for structure, response to lack of structure, knowing style, creating style and planning style. All individual traits, except for two, had no significant association with IRS length chosen. The two that did were the CoSI Knowing Style and Creating Style. Knowing Style was significantly correlated to IRS length,  $r = .1$ ,  $n = 538$ ,  $p < .05$ , using a two-tailed test. However, this positive correlation is very weak at .1, and so one cannot be conclusive about an association between someone's desire to 'know' (in terms of their cognitive style) and their propensity to define IRSs with a greater number of categories. Creating Style was significantly correlated to IRS length,  $r = .122$ ,  $n =$

538,  $p < .01$ , using a two-tailed test. This positive correlation, whilst slightly stronger than for Knowing Style, is also very weak at .122, and so it was also not possible to be conclusive about an association between someone's desire to 'create' (in terms of their cognitive style) and their propensity to define IRSs with a greater number of categories.

On the whole, conclusive associations between all individual characteristics and IRS length were not found.

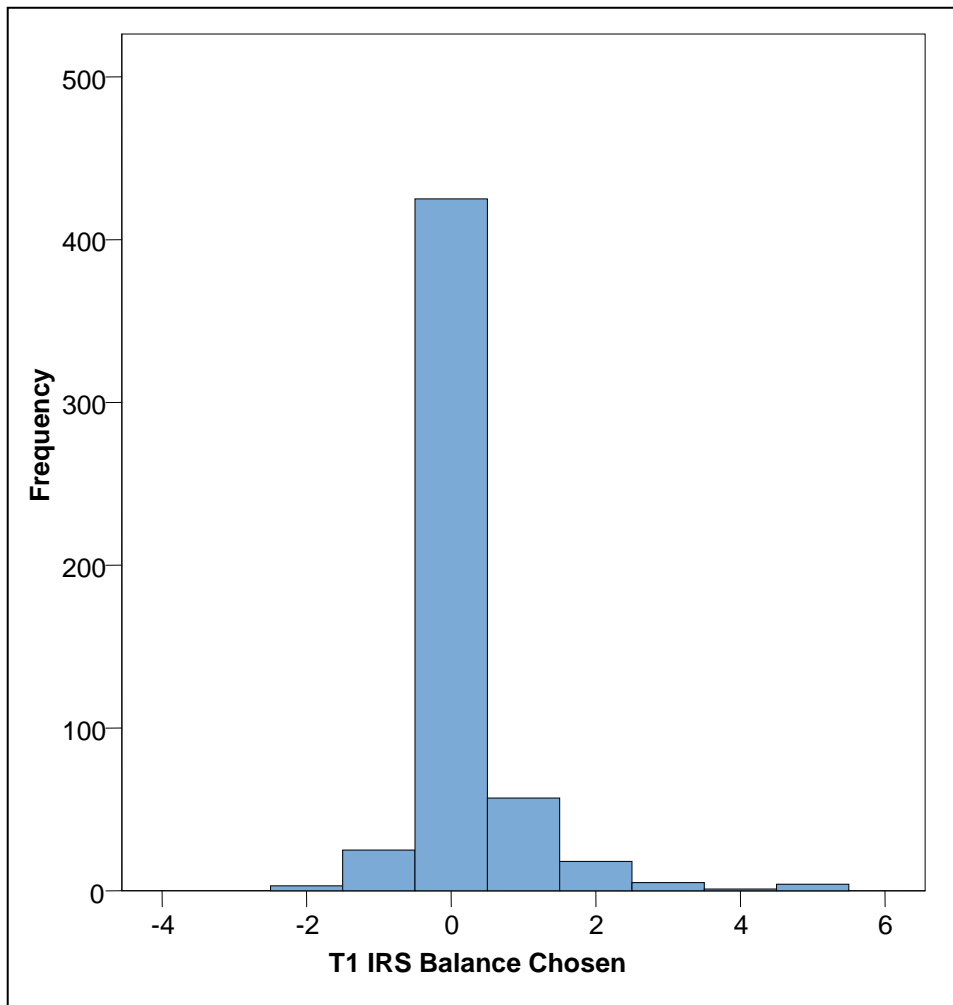
### 5.5.3 IRS Balance Chosen

Insights from the qualitative phase showed that some respondents genuinely appeared to require a greater number of categories for one side of the agreement/disagreement bipolar continuum than for the other. Given bipolar fixed rating-scales in survey research are virtually always balanced or symmetrical (i.e. the same number of categories on both sides of the continuum), it was informative to see whether respondents actually *opted* for a balanced IRS when given the choice.

As with IRS length, the survey system captured respondents' choice of categories for each of the endpoints both *before* the practice (on the sixteen uncorrelated items), and *after*, should the respondent have opted to modify their categories. This means that in any one time period the term IRSB\_1 is used when referring to the balance of respondents' first IRS, and the term IRSB\_2 is used when referring to their second (if applicable). The term IRSB\_Used refers to the balance of the IRS that the respondent used when completing the main survey; which would be IRSB\_1 for those who opted out of changing their IRS, and IRSB\_2 for those who chose to modify it. If an IRS is balanced, and therefore has the same number of categories on either side of the neutral

point, it would have an IRS balance score of 0. If the IRS has a greater number of categories on the (dis)agreement side of the continuum, then the score would be positive (negative). The score for imbalanced IRSs (i.e. negative or positive) represents the number of extra categories one pole has over the other. For example, an IRS balance score of 3 would indicate that the respondent's IRS had three extra categories on the agreement side of the continuum.

Figure 5. 18 illustrates the spread of the IRSB\_Used for respondents (the rating-scale balance ultimately chosen) in a histogram. The mean IRSB\_Used was .19 with a standard deviation of .757. It is quite clear from the graph that the mode rating-scale balance was 0, with 425 opting for a perfectly symmetrical IRS. Overall, it is clear that most respondents, even when given the choice, prefer to have the same number of categories for both agreement and disagreement. The most imbalanced IRS defined had five extra categories on the agreement side of the continuum, chosen by four respondents. The most imbalanced IRS in the opposite direction was two extra categories for the disagreement side of the continuum, chosen by three respondents. It would seem that, where respondents preferred imbalanced IRSs, they tended to be skewed towards having extra categories for agreement. Of these respondents, 83% defined balanced IRSs (i.e. a score of 0). Nevertheless it appears that approximately one in five people prefer to have an imbalanced rating-scale.



**Figure 5. 18 T1 IRS Balance Used**

Out of the 538 who completed the IRSP in T1, the 116 who chose to modify their IRS were examined in the last section (section 5.5.2) for changes to their IRS lengths. It was established that there was a difference between IRSL\_1 and IRSL\_2 (in T1). In other words, these respondents had modified their rating-scale lengths before proceeding onto the main survey. However, while the lengths may have changed, the balance may have stayed the same. For example, if a respondent were to define an IRS of  $-3 \leftarrow 0 \rightarrow 3$ , it would have a length of 7 categories and a balance score of 0. Should the respondent modify it to  $-4 \leftarrow 0 \rightarrow 4$ , it would now have a length of 9 categories but it would still have a balance score of 0. For this reason, a paired sample t-test was conducted on both IRS balance scores for those 116 who opted to modify their IRS. The results showed that

respondents' IRSB\_1 in T1 (mean = .03; SD = .665) were significantly different from their IRSB\_2 in T1 (mean = .22; SD = .759);  $t(df = 115) = -2.553$ ,  $p = .012$ , CI = -.352 to -.044. This difference, although statistically significant, was quite small. It suggested that for those who chose to modify their IRS in T1, some chose to add additional categories on the agreement side of their IRS. However, this difference is somewhat too small to be of substantial importance.

Of those who completed the IRSP in T1, 192 of those invited, returned to complete the IRSP in T2. Of the 116 respondents who modified their IRS in T1, 46 completed the IRSP survey in T2. A paired sample t-test was conducted on those 46 respondents to see whether their IRSB\_2 (effectively their IRSB\_Used) from T1 was significantly different to their IRSB\_1 in T2. In other words, this test was undertaken to see whether the modified version of their IRS in T1 was replicated the first time they defined their IRS some time later in T2. The results showed that respondents' IRSB\_2 in T1 (mean = .28; SD = .750) were not significantly different from their IRSB\_1 in T2 (mean = .15; SD = .363);  $t(df = 45) = 1.182$ ,  $p = .244$ , CI = -.092 to .353. This supports the previous finding which examined IRSL\_2 in T1 and IRSL\_1 in T2, where no significant difference was found in the rating-scale lengths. This stability, between IRSB\_2 in T1 and IRSB\_1 in T2, would also suggest that the balance of their modified IRS in T1 (IRSB\_2) was personally appropriate, in that these respondents chose to have the very same IRS balance the next time they completed the survey. In fact, for those who modified their IRS in T1, a paired sample t-test found no significant difference between the IRSB\_Used in T1 (mean = .28; SD = .750) and the final IRSB\_Used in T2 (mean = .20; SD = .453);  $t(df = 45) = .752$ ,  $p = .456$ , CI = -.146 to .320. This addressed the possibility that respondents could have modified the balance of their IRSs again in T2.

This supports the earlier theory that once respondents adopted their modified IRS in T1, it appeared to be stable even after a wash-out period.

As for those who did not choose to modify their IRS balance in T1, there was a significant difference between their IRSB\_Used in T1 (mean = .15; SD = .579), and their IRSB\_Used in T2 (mean = .02; SD = .398);  $t(df = 145) = 2.613$ ,  $p = .010$ , CI = .032 to .229. This difference, although statistically significant, is too small to be of substantial importance. It might suggest that some respondents who did not modify their IRS in T1, decided to have more balanced IRSs in T2, although overall IRS lengths did not change (as established in section 5.5.2 on IRS Lengths Chosen).

It was worth comparing mood scores for the 192 respondents who completed the IRSP in T1 and in T2, to see if there were any differences in respondents' temporary states. A paired sample t-test showed that there was no significant difference between the scores on the mood measure in T1 (mean = .228; SD = .499) and the mood measure in T2 (mean = .278; SD = .495);  $t(df = 191) = -1.113$ ,  $p = .267$ , CI = -.136 to .038. This meant that it was impossible to test whether the stability of both IRS length and balance is affected by temporary states like mood, given the mood of respondents was, on average, similar in both time periods.

#### *5.5.3.1 Individual Characteristics and IRS Balance*

In an earlier section (on page 5.72), it was established that there were no observable relationships between respondents' individual characteristics and their choice of IRS length. However, there may have been a relationship between these individual

characteristics and respondents' propensity to choose a balanced or imbalanced IRS. As such, this was also examined.

The IRS balance scores were rank-ordered and the Mann-Whitney U-test was used to compare the ranks for the males (n=164) versus the females (n=374). The results indicated no significant difference between the genders,  $U = 29583$ ,  $p=.358$ , with the sum of the ranks equal to 43113 for males and 101878 for females.

Next the Mann-Whitney U-test was used to compare the ranks for the undergraduates (n=421) versus the postgraduates (n=117). The results indicated no significant difference between the two groups,  $U = 24196$ ,  $p=.683$ , with the sum of the ranks equal to 113027 for undergraduates and 31964 for postgraduates.

Finally the Kruskal-Wallis test was used to evaluate differences among the ten categories of degree area, shown in Table 5. 46. The outcome of the test indicated a significant difference among the groups,  $H = 20.596$  ( $df=9$ ,  $N=535$ ),  $p=.015$ . This result suggests that at least one of the groups has a different propensity to choose balanced/imbalanced IRSs than the others. This result highlights an area for further investigation.

**Table 5. 46 T1 IRSP: Kruskal-Wallis Rankings for IRS Balance by Degree Area.**

Degree_Area	N	Mean Rank
Arts	10	323.25
Business	38	316.82
Engineering	18	307.56
History	69	253.90
IT	16	241.00
Languages	111	263.50
Law	50	278.48
Life Sciences	116	251.02
Physical Sciences	28	241.14
Social Sciences	79	280.44
Total	535*	

\* 3 respondents' degree areas were unknown which is why n=535 and not 538.

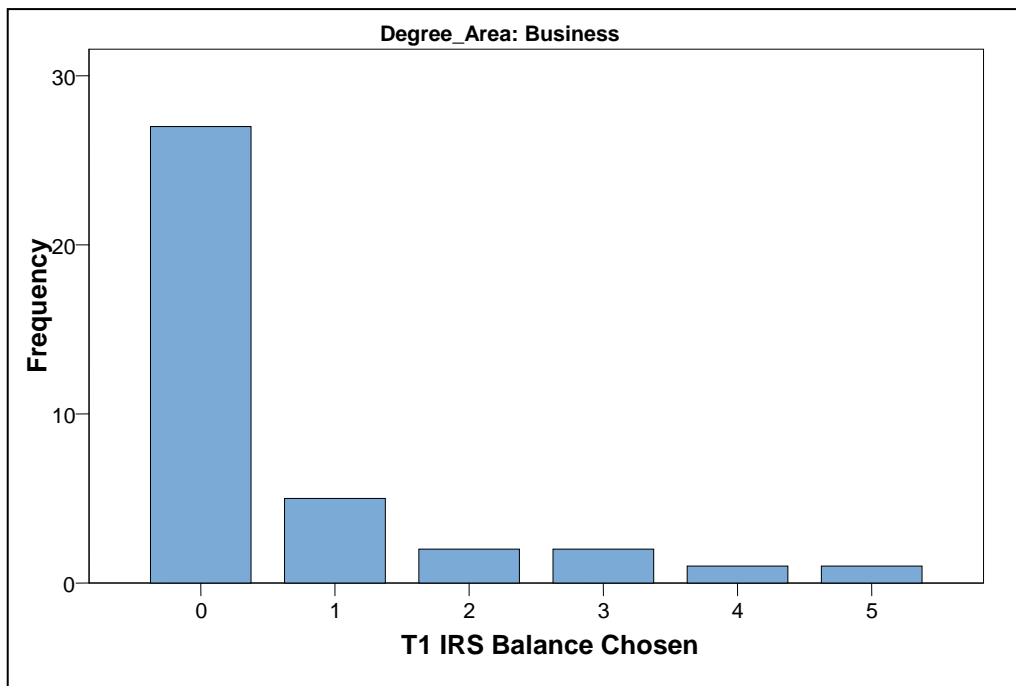
Given the above result, it was useful to see what the mean IRS balance score was for each group. Table 5. 47 shows the mean balance scores for each degree area grouping. It was very interesting to see that groups who chose perfectly balanced IRSs (i.e. had a mean IRS balance of 0) were those respondents from the 'Physical Sciences' and 'IT' disciplines. These groups also had a very narrow range of IRS balance scores, from -1 to 1, and similarly small standard deviations (.385 and .365 respectively). This is in contrast to some of the other groups. For example respondents from the Business discipline had a mean IRS balance of .63 (SD = 1.239), which is the most imbalanced of the mean scores. The scores for this group ranged from 0 to as much as 5. It would seem that students from this discipline had a greater tendency to choose a larger number of gradations for the agreement side of their continuum. Whilst respondents from the Arts discipline also scored a high mean of .6, this group is too small to draw comparisons against some of the other groups, as it consists of only 10 respondents.



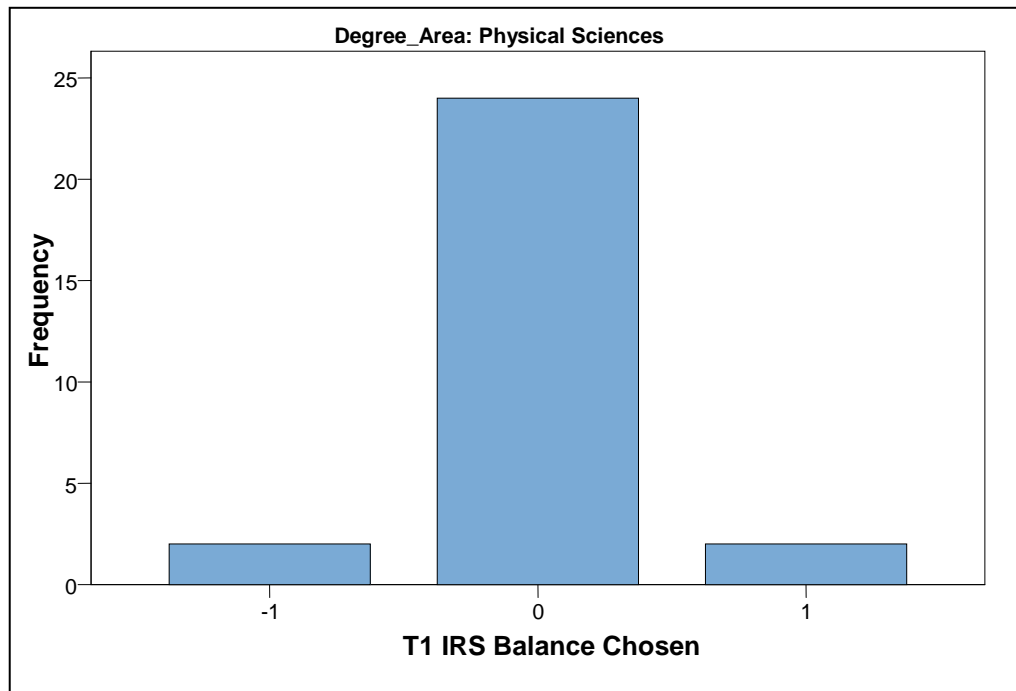
**Table 5. 47 T1 IRSP: Mean Respondent IRS Balance Scores Grouped by Degree Area.**

Degree_Area		N	Minimum	Maximum	Mean	Std. Dev.
Arts	T1_Numer_Balance_Used	10	0	2	.60	.966
Business	T1_Numer_Balance_Used	38	0	5	.63	1.239
Engineering	T1_Numer_Balance_Used	18	-1	1	.28	.575
History	T1_Numer_Balance_Used	69	-2	3	.12	.738
IT	T1_Numer_Balance_Used	16	-1	1	.00	.365
Languages	T1_Numer_Balance_Used	111	-1	2	.12	.518
Law	T1_Numer_Balance_Used	50	-2	5	.28	1.051
Life Sciences	T1_Numer_Balance_Used	116	-2	5	.09	.710
Physical Sciences	T1_Numer_Balance_Used	28	-1	1	.00	.385
Social Sciences	T1_Numer_Balance_Used	79	0	2	.19	.455
Total		535				

Figure 5. 19 and Figure 5. 20 illustrate the dissimilar distributions of the IRS balance scores for the respondents from the ‘Business’ and the ‘Physical Sciences’ disciplines.



**Figure 5. 19 T1 IRSP: Distribution of IRS Balance Scores for Respondents within the Business Discipline.**



**Figure 5. 20 T1 IRSP: Distribution of IRS Balance Scores for Respondents within the Physical Sciences Discipline.**

Next, individual traits were examined for any associations with IRS balance. Spearman correlations were obtained between IRSB\_Used in T1 and each of the individual trait scores from T1; extraversion, conscientiousness, agreeableness, neuroticism, openness, desire for structure, response to lack of structure, knowing style, creating style and planning style. The Spearman correlations indicated that all the individual traits had no significant association with IRS balance chosen.

Given differences were found between respondents studying for different degrees, the data was split by degree area and the Spearman correlations between individual traits and IRS balance were recalculated. It is duly noted that some of the groups contained too small a sample size to have confidence in the outcome of the test. As such, significant correlations, between the variables found within small groups, are ignored. Interestingly, for the respondents from a Business discipline ( $n = 38$ ), Affective Orientation was positively correlated with IRS balance,  $r = .322$ ,  $p < .05$ , two-tail test. This meant that for those studying business degrees, the higher their affective

orientation, the greater their propensity to choose an imbalanced IRS with more categories on the agreement side of the continuum, on average. However, none of the other degree area groupings yielded an association between these two variables, see Table 5. 48. This was the only statistically significant association between individual traits and IRS balance that was identified when the sample was split by degree area.

Overall, conclusive associations between all individual characteristics and IRS balance were not found. When the data was split by degree area, an association between Affective Orientation and IRS balance was found for those who studied Business.

**Table 5. 48 T1 IRSP: Spearman’s rho Correlations between AO and IRS Balance, by Degree Area.**

Degree_Area			T1 IRS Balance	T1 AO Score
Arts	T1_Numer_Balance_Used	Correlation Coefficient	1.000	.190
		Sig. (2-tailed)	.	.599
		N	10	10
<b>Business</b>	<b>T1_Numer_Balance_Used</b>	<b>Correlation Coefficient</b>	<b>1.000</b>	<b>.322*</b>
		<b>Sig. (2-tailed)</b>	.	<b>.049</b>
		<b>N</b>	<b>38</b>	<b>38</b>
Engineering	T1_Numer_Balance_Used	Correlation Coefficient	1.000	.001
		Sig. (2-tailed)	.	.996
		N	18	18
History	T1_Numer_Balance_Used	Correlation Coefficient	1.000	-.105
		Sig. (2-tailed)	.	.388
		N	69	69
IT	T1_Numer_Balance_Used	Correlation Coefficient	1.000	-.038
		Sig. (2-tailed)	.	.888
		N	16	16
Languages	T1_Numer_Balance_Used	Correlation Coefficient	1.000	.117
		Sig. (2-tailed)	.	.221
		N	111	111
Law	T1_Numer_Balance_Used	Correlation Coefficient	1.000	.036
		Sig. (2-tailed)	.	.803
		N	50	50
Life Sciences	T1_Numer_Balance_Used	Correlation Coefficient	1.000	-.031
		Sig. (2-tailed)	.	.737
		N	116	116
Physical Sciences	T1_Numer_Balance_Used	Correlation Coefficient	1.000	.164
		Sig. (2-tailed)	.	.405
		N	28	28
Social Sciences	T1_Numer_Balance_Used	Correlation Coefficient	1.000	.016
		Sig. (2-tailed)	.	.890
		N	79	79

\* Correlation is significant at the 0.05 level (2-tailed).

### 5.5.4 IRS Verbal Labels Chosen

In the literature review, the issue of verbal labels was explored. On the issue of verbal anchoring, the objective of the IRSP was to help the respondent to access what, for them, are meaningful verbal labels for both endpoints of their continuum. The verbal labels chosen are meant to represent a respondent’s conceptual extreme for

agreement/disagreement. The verbal labels chosen by respondents were examined, so as to give an indication of whether or not they appeared to be verbally anchoring their IRSs properly (i.e., selecting appropriate adverbs to indicate an extreme position). So too, was it interesting to see what verbal labels were actually chosen and *which* proved to be the most popular.

Before examining the verbal labels, the data had to be cleaned. Spelling mistakes were corrected so that if a respondent, for example, intended to use ‘Definitely’ and wrote ‘Definately’, his/her true intention was recorded.

Table 5. 49 shows the verbal labels that were chosen by those respondents who completed the IRSP in time period 1,  $n = 538$  (British, English-speaking group only). Some of the labels have been struck through, as they were deemed to be inappropriate, either because the respondent had chosen a *synonym* for ‘agree’ (which is not what was intended by the IRSP instructions), or because they had clearly misunderstood the instructions in some other way. For the ‘agree’ endpoint, 6 out of the 538 labels were deemed inappropriate (1% error rate).

The top five most popular verbal labels in T1 for ‘agree’ (the extreme position) were ‘Completely’ (28.4%), ‘Totally’ (27.5%), ‘Definitely’ (11.5%), ‘Absolutely’ (10%) and ‘Strongly’ (7.8%). These verbal labels were also the same top five chosen in T2: Of the 328 that completed the IRSP in T2 the top five chosen were ‘Completely’ (36.3%), ‘Totally’ (23.2%), ‘Strongly’ (11.6%), ‘Definitely’ (11.0%), and ‘Absolutely’ (7.9%).

**Table 5. 49 T1 IRSP: Verbal Labels Used for Agreement Endpoint.**

	Frequency	Percent
Completely	153	28.4
Totally	148	27.5
Definitely	62	11.5
Absolutely	54	10.0
Strongly	42	7.8
Really	16	3.0
one hundred percent	11	2.0
Fully	8	1.5
Entirely	7	1.3
Wholeheartedly	7	1.3
Very much	4	.7
<del>Agree</del>	<del>2</del>	<del>.4</del>
Categorically	2	.4
Certainly	2	.4
Highly	2	.4
Utterly	2	.4
Very strongly	2	.4
Amazingly	1	.2
Comprehensively	1	.2
<del>don't really</del>	<del>4</del>	<del>.7</del>
<del>I do</del>	<del>4</del>	<del>.7</del>
<del>I tend to</del>	<del>4</del>	<del>.7</del>
Literally	1	.2
Most assuredly	1	.2
Mostly	1	.2
Positively	1	.2
Seriously	1	.2
Undoubtedly	1	.2
Unequivocally	1	.2
Wholly	1	.2
<del>Yes</del>	<del>4</del>	<del>.7</del>
Total	538	100.0

Those verbal labels that have been struck through were deemed to be inappropriate labels.

The inspection process was repeated for the ‘disagree’ verbal labels. Table 5. 50 shows the verbal labels chosen in T1 for the disagreement side of the continuum. As was done previously, the inappropriate labels were noted. For ‘disagree’, 12 out of the 538 labels were deemed to be inappropriate. This is still quite a low error rate at 2%, but might indicate there is room for improvement in the IRSP instructions to further reduce it.

**Table 5. 50 T1 IRSP: Verbal Labels Used for Disagreement Endpoint.**

	Frequency	Percent
Completely	203	37.7
Totally	133	24.7
Absolutely	44	8.2
Strongly	43	8.0
Really	27	5.0
Definitely	17	3.2
<del>Don't</del>	8	1.5
one hundred percent	7	1.3
Utterly	7	1.3
Entirely	6	1.1
Highly	4	.7
Vehemently	4	.7
<del>No way</del>	3	.6
Very much	3	.6
Certainly	2	.4
Fully	2	.4
Most definitely	2	.4
Positively	2	.4
Seriously	2	.4
Very strongly	2	.4
Wholeheartedly	2	.4
<del>Also</del>	4	.2
<del>Always</del>	4	.2
Categorically	1	.2
<del>Couldn't</del>	4	.2
<del>Disagree</del>	4	.2
f*****	1	.2
Fundamentally	1	.2
Honestly	1	.2
Hugely	1	.2
Literally	1	.2
Massively	1	.2
Mostly	1	.2
Respectfully	1	.2
So	1	.2
Thoroughly	1	.2
Total	538	100.0

Those verbal labels that have been struck through were deemed to be inappropriate labels.

The top five most popular verbal labels for 'disagree' (the extreme position) were 'Completely' (37.7%), 'Totally' (24.7%), 'Absolutely' (8.2%), 'Strongly' (8.0%) and 'Really' (5%). These verbal labels were similarly popular in T2: Of the 328 that

completed the IRSP in T2 the top five chosen were ‘Completely’ (44.2%), ‘Totally’ (21.0%), ‘Strongly’ (11.0%), ‘Definitely’ (4.6%), and ‘Absolutely’ (4.6%).

A noteworthy result is that, in T1, the adverb ‘Strongly’ was 5<sup>th</sup> and 4<sup>th</sup> most popular for ‘agree’ and ‘disagree’ respectively, with approximately only 8% of respondents choosing it for either of the endpoints. Given most standardised Likert-type rating-scales (LTRSs) use ‘strongly agree’ and ‘strongly disagree’ as verbal anchors for endpoints, this result is most insightful.

The question of verbal label symmetry was an area worth exploring. In the qualitative phase of research, it was apparent that some respondents felt that their extreme agree/disagree endpoints required different verbal labels. As such, the survey data was examined to see whether this preference was replicated for any of the respondents. Table 5. 51 shows the number of respondents in T1 who chose verbal labels that were the same for both endpoints, and those where the endpoints were different. The majority, at 58.7%, opted for different verbal labels for their endpoints. The verbal labels chosen by the 192 out of the 538 that completed the IRSP in both time periods are shown in Table 5. 52. This shows that within the very same group of respondents, 53.5% of them chose different verbal labels for their endpoints in T1, but only 39.6% of them did so in T2. This indicates that some may have opted to change from imbalanced to balanced verbal labels the second time around.



**Table 5. 51 T1 IRSP: Verbal Label Symmetry.**

	Frequency	Percent
Same	222	41.3
Different	316	58.7
Total	538	100.0

**Table 5. 52 T1-T2: IRSP-IRSP Test Group, Verbal Label Symmetry.**

	T1 Frequency	T2 Frequency	T1 Percent	T2 Percent
Same	89	116	46.4	60.4
Different	103	76	53.6	39.6
Total	192	192	100.0	100.0

Of the 116 respondents in T1 who opted to modify their IRS, only 1.7% modified their ‘agree’ verbal labels, and 0.9% modified their ‘disagree’ verbal labels. This indicated that on the whole, where respondents chose to modify their IRSs, they were doing so in order to modify the *lengths* of their rating-scales (i.e. number of categories) as opposed to their verbal anchoring.

**5.5.5 LTRS versus IRS: Respondent Preferences**

In time period 2, at the end of the survey, respondents (who responded to at least one IRSP survey) were asked a set of questions pertaining to the attention given to survey questions (Attention), meaningfulness of the ratings (Meaningful), preference of rating-scale (Preference), and the ease of designing an IRS. The four questions posed to those who were in Test Groups 2 (TG2) and 3 (TG3) (i.e. IRSP-LTRS, LTRS-IRSP), are shown in Figure 5. 21:

- |  |
|--|
| <p>[ATTENTION]</p> <ul style="list-style-type: none"> <li>• I pay more attention when answering survey questions if I use my OWN SCALE rather than a FIXED SCALE.</li> </ul> <p>[MEANINGFUL]</p> <ul style="list-style-type: none"> <li>• My answers more accurately reflect my opinions if I use my OWN SCALE rather than a FIXED SCALE.</li> </ul> <p>[PREFERENCE]</p> <ul style="list-style-type: none"> <li>• I would prefer to design and use my OWN SCALE rather than a FIXED SCALE, when answering survey questions.</li> </ul> <p>[EASE]</p> <ul style="list-style-type: none"> <li>• I found it easy to design my OWN SCALE.</li> </ul> |
|--|

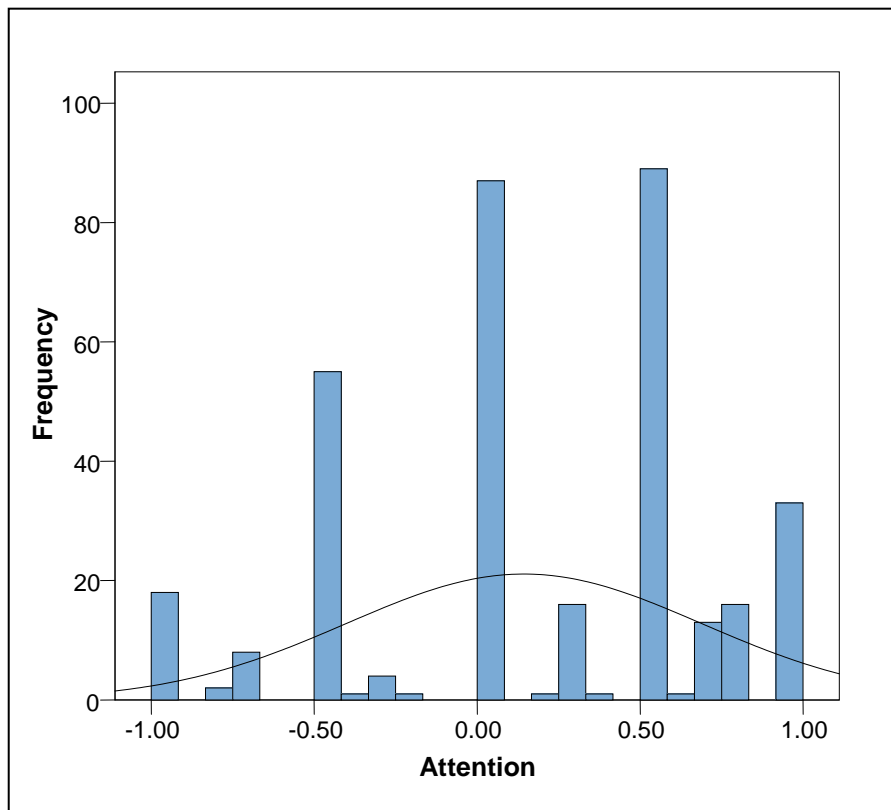
**Figure 5. 21 IRSP feedback questions posed to TG2 and TG3 at end of survey.**

A precursor to these questions was an instruction page explaining to respondents what is meant by ‘fixed scale’ and by ‘own scale’. The total number of respondents who experienced both methods were 210 in TG2 (IRSP-LTRS) and 136 in TG3 (LTRS-IRSP). Respondents in time period 2 will have used the respective method of that survey to rate the feedback questions. So, respondents in TG2 will have rated their answers to the feedback questions using a fixed LTRS of  $-2 \leftarrow 0 \rightarrow 2$ , verbally anchored with ‘strongly disagree’, ‘neutral’, and ‘strongly agree’, whereas, respondents in TG3 will have used their IRS to rate the feedback questions. As with all the analyses, all the ratings were indexed from -1 to 1.

An inspection of the mean values for both TG2 and TG3, suggested that respondents had a positive experience in using the IRSP (Table 5. 53). The mean values for all four items are positive and greater than 0 (the neutral position). Figure 5. 22 through to Figure 5. 25 show the spread of responses to these four items. Note that the bars peak at  $\pm 1$ ,  $\pm 0.5$  and 0, due to those using the fixed intervals of the LTRS. The curves seem marginally skewed in favour of the IRSP over LTRS. However, further testing was needed to confirm this.

**Table 5. 53 Feedback on the use of the IRSP over LTRS: Test Groups 2 and 3**

		Attention	Meaningful	Preference	Ease
N	Valid	346	346	346	346
	Missing	0	0	0	0
Mean		.1438	.2791	.1775	.3521
Std. Deviation		.54531	.52532	.55112	.57666
Minimum		-1.00	-1.00	-1.00	-1.00
Maximum		1.00	1.00	1.00	1.00



**Figure 5. 22 Test Groups 2 & 3: Spread of Responses to 'Attention' Item.**

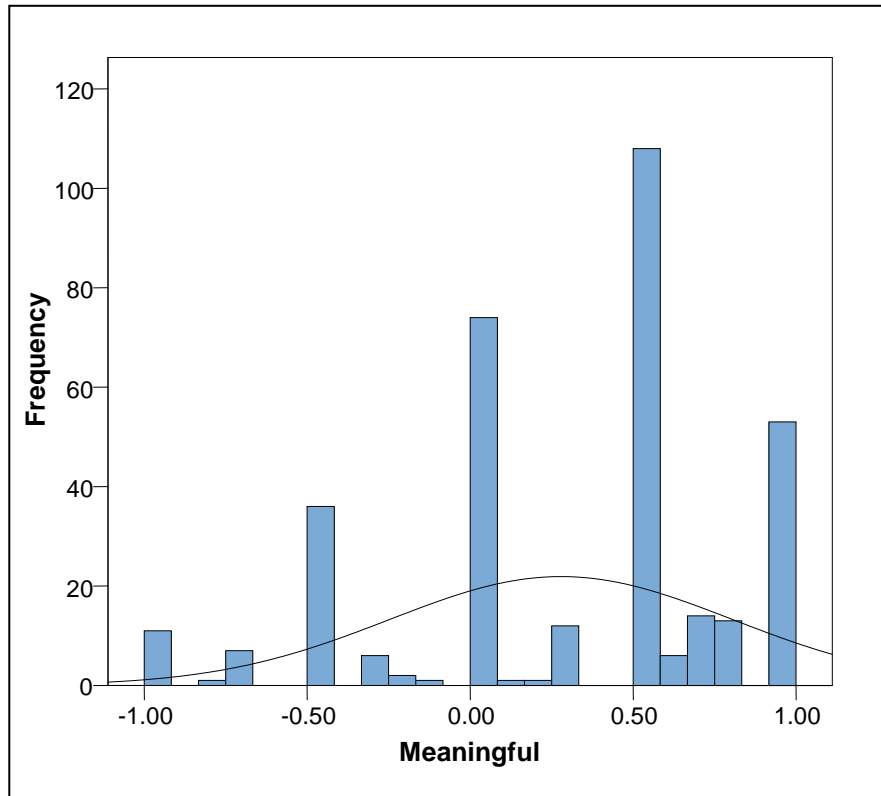


Figure 5. 23 Test Groups 2 & 3: Spread of Responses to 'Meaningful' Item.

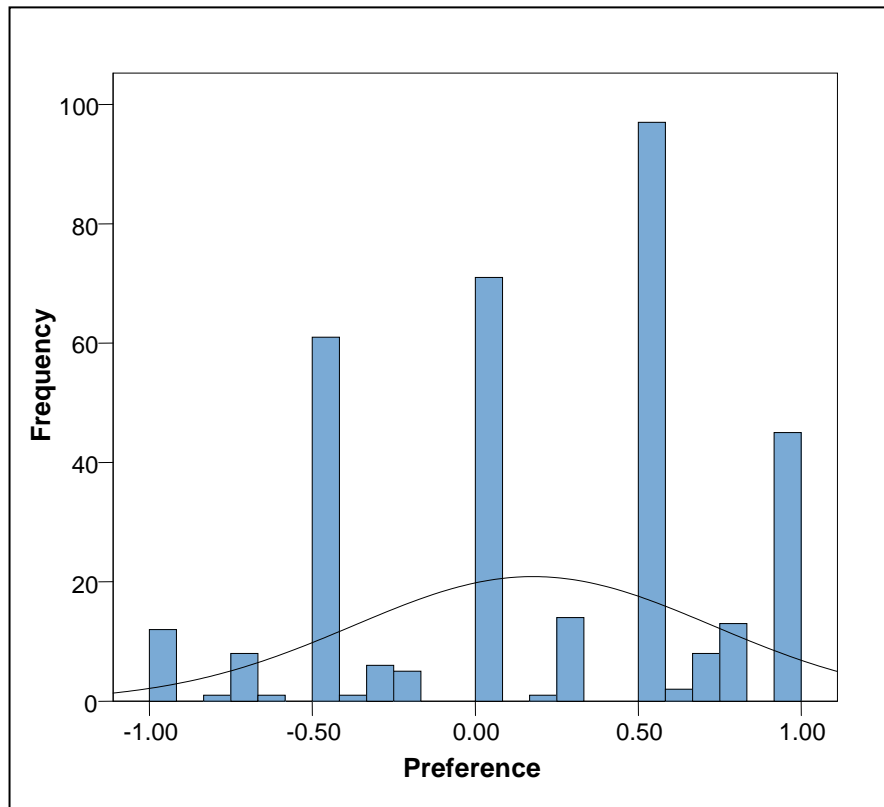


Figure 5. 24 Test Groups 2 & 3: Spread of Responses to 'Preference' Item.

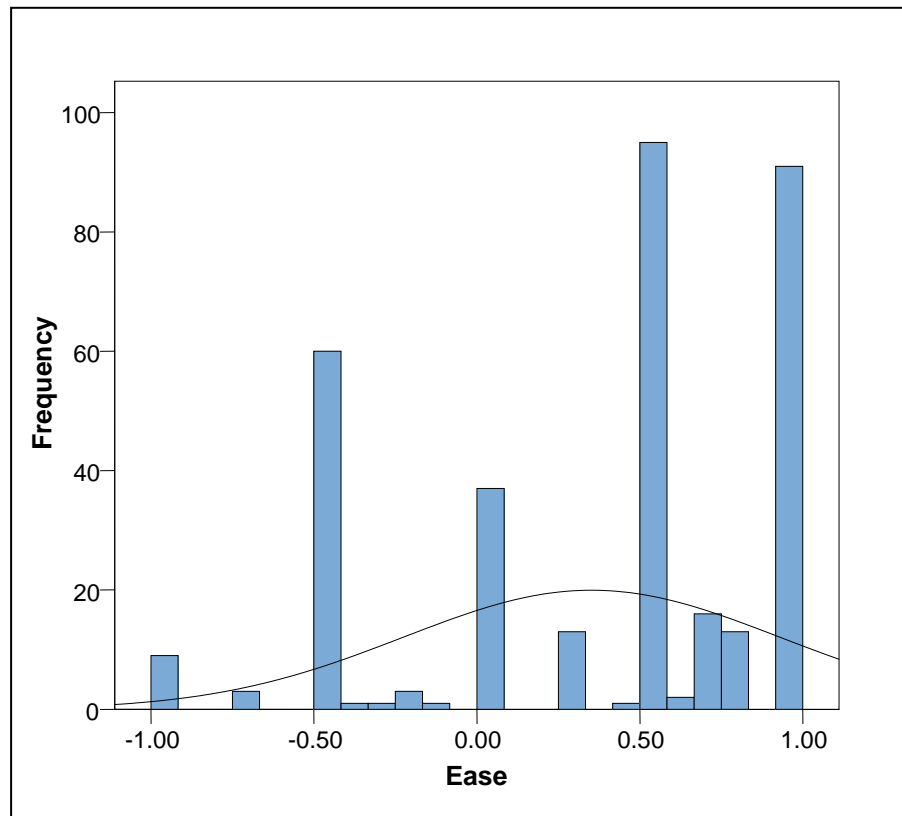


Figure 5. 25 Test Groups 2 & 3: Spread of Responses to 'Ease' Item.

The data was split by Test Group, given there may have been differences between the two groups in their assessment of the IRSP. A one-sample t test was conducted<sup>12</sup> to test the null hypothesis, that respondents would have a mean neutral opinion on all four items (Table 5. 54). The values for all four items, in both Test Groups, are significantly below 0.05 (in fact, all but one are significant at the .001 level), which indicates that the mean ratings are significantly different from 0. Additionally, none of the confidence intervals for the mean differences contains zero, which also indicates that the difference is significant.

<sup>12</sup> From the central limit theorem, one-sample t tests are robust to violations of the normality assumption as long as n is large (Gravetter and Wallnau 2004), which it is here.

**Table 5. 54 Test Groups 2 & 3: One-sample t test on IRSP feedback items.**

Test Group		Test Value = 0					
		t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
						Lower	Upper
TG2 IRSP-Likert	Attention	2.458	209	.015	.08095	.0160	.1459
	Meaningful	7.871	209	.000	.26190	.1963	.3275
	Preference	3.724	209	.000	.13333	.0627	.2039
	Ease	4.287	209	.000	.16905	.0913	.2468
TG3 Likert-IRSP	Attention	4.491	135	.000	.24096	.1348	.3471
	Meaningful	6.078	135	.000	.30574	.2063	.4052
	Preference	4.834	135	.000	.24581	.1452	.3464
	Ease	16.095	135	.000	.63485	.5568	.7129

The findings of the one-sample t-tests indicate that for both Test Groups, the IRSP was preferred over the LTRS on the three feedback items where respondents were asked to make a direct comparison (Attention, Meaningful, and Preference). ‘Ease’ required respondents to rate the level of ease they experienced when designing their IRS, and the results clearly show that the sample generally found it easy. On inspection of TG2, the IRSP was particularly favoured over the LTRS when it came to the ‘meaningfulness’ of rating-scales to respondents, scoring a mean rating of .26. This meant that, on average, the respondents felt that the IRS was more capable of accurately reflecting their opinions than the LTRS. TG3 also produced a similar result for item ‘Meaningful’, with a mean rating of .31. However, it is also clear that the mean ratings for TG3 were consistently higher than those for TG2. This is particularly obvious when examining the mean ratings for ‘Ease’ in both groups; TG2 (.17) and TG3 (.63). Whilst the sample size for TG2 is larger than that for TG3, they are both adequately large. This difference is likely to be the result of an order-effect (i.e. which method was used first).

To investigate this further, a one-way ANOVA was used (Table 5. 55). Both ‘Ease’ and ‘Attention’ had significance values below .05. This means that there were significant differences between TG2 and TG3 on how respondents rated the IRSP for ‘Ease’ and ‘Attention’. This suggested that there may have been an order-effect which impacted on the mean ratings of ‘Attention’ and ‘Ease’.

**Table 5. 55 Test Groups 2 & 3: One-way ANOVA to test for order-effects on the mean ratings of IRSP feedback items.**

		Sum of Squares	df	Mean Square	F	Sig.
Attention	Between Groups	2.113	1	2.113	7.235	.007
	Within Groups	100.478	344	.292		
	Total	102.591	345			
Meaningful	Between Groups	.159	1	.159	.574	.449
	Within Groups	95.049	344	.276		
	Total	95.208	345			
Preference	Between Groups	1.044	1	1.044	3.463	.064
	Within Groups	103.744	344	.302		
	Total	104.788	345			
Ease	Between Groups	17.910	1	17.910	63.637	.000
	Within Groups	96.814	344	.281		
	Total	114.724	345			

### 5.5.6 Summary

This chapter presented a discussion of the findings pertaining to the *testing* of the IRSP. In Stage 1, confirmatory factor analysis (CFA) was applied to the data split by method (IRSP vs LTRS), to establish loose cross-validation for three of the measurement models; Affective Orientation (AO), Personal Need for Structure (PNS), and Cognitive Style Indicator (CoSI). The fit indices for the AO measurement model indicated a less than adequate fit for both groups (IRSP and LTRS). The fit statistics were extremely similar for both groups, with both falling slightly short of the ideal cut-off values. Convergent validity of the model was present for both groups, yet the model fit

demanded improvements. Model diagnostics were examined to improve model fit. The re-specified AO model (six out of the fifteen items were removed), yielded acceptable fit statistics for both groups, with a marginally better fit for the LTRS.

Stage 1 also showed that loose cross-validation was present across groups with the Personal Need for Structure (PNS) model. The fit statistics indicated a similar fit for both groups, and was very close to being acceptable (all except for the RMSEA value, which was slightly inadequate in both groups). After three weaker items were removed, the re-specified PNS model yielded an acceptable fit for both groups, with a marginally stronger fit for the IRSP group. With the Cognitive Style Indicator (CoSI) measurement model, both groups demonstrated inadequate fit, although the IRSP performed marginally better than the LTRS. On the whole, factor loadings were higher for the IRSP group than they were for the LTRS. Convergent validity, although substandard for two out of the three factors, was marginally better with the IRSP. There was an adequate level of discriminant validity for both groups, although the IRSP performed particularly better than the LTRS here. After the removal of four weaker items, the re-specified CoSI model yielded an acceptable fit for both groups, with similar fit statistics. These results indicate that the IRSP, as an instrument, is equally capable of capturing data that replicates the psychometric measurement models obtained through the pre-validated LTRS methods. Where the LTRS data yielded a fit, so too did the IRSP data. Similarly, where the LTRS data yielded a problem with either the fit or particular items, the IRSP did also.

Stage 2 applied further cross-validation testing across the two groups, using the re-specified models obtained from Stage 1. This was done through further CFA analyses to



test for measurement equivalence, through tighter cross-validation procedures. The IRSP data and the LTRS data passed the test of tight cross-validation on the AO measurement model, demonstrating factor structure, factor loading and error variances equivalence. Both groups also possessed strong cross-validation for the PNS and CoSI models, demonstrating factor structure, factor loading, and inter-factor covariance equivalence in each. This indicated that the IRSP and the LTRS data yielded the same underlying measurement model, for each of the three constructs. Ideally, scalar equivalence would also have been tested, to establish whether the numerical ratings on each of the rating-scales (IRSP vs LTRS) ‘meant’ the same thing across groups. However this test was not possible under ADF estimation.

Stage 3 assessed test-retest reliability for both groups, by examining T1 to T2 correlations on all the psychometric constructs, for TG1 (IRSP-IRSP) and TG4 (LTRS-LTRS). Results suggested that both methods performed similarly well. They did not indicate that one method was performing better than the other. Furthermore, an additional test of construct validity was conducted, extending the initial tests in Stage 1. Here, all the psychometric factors were examined together, using an approach similar to Campbell and Fiske’s MTMM matrix. The results confirmed a finding from Stage 1, that reliability (internal consistency) was present for the AO, PNS and CoSI factors (with both IRSP and LTRS groups), but the Big Five Inventory (BFI) factors did not possess adequate internal consistency levels with either of the groups. The AO, CoSI and PNS factors achieved both convergent and discriminant validity under Campbell and Fiske’s MTMM method.

Stage 4 examined respondents IRS choices (numerical endpoints, degree of balance, and verbal labels). A seven-point IRS was the most popular IRS length chosen. The facility for respondents to practice their IRS on sixteen uncorrelated items followed by the option to modify it, proved worthwhile. This was evidenced by the fact that around one fifth of respondents (in T1) chose to modify their IRSs through this facility. Paired sample t-tests showed that these respondents kept the same IRSs when completing the survey in T2. On the whole, those that *did not* opt to modify their IRSs within the T1 survey, kept the very same IRSs when completing the survey in T2. There were no observable relationships between respondents' demographic characteristics (gender, level of study, and degree discipline) and their choice of IRS length. Whilst, there were two (out of the eleven) personal traits (factors) that were significantly positively correlated with IRS length, the correlations were very weak. As such, conclusive associations between individual characteristics and IRS length were not found.

Four fifths of respondents opted for a balanced IRS (i.e. with the same number of intervals on either side of neutral). Where respondents preferred imbalanced rating-scales it tended to be in favour of having extra categories on the 'agreement' side of the continuum. For respondents who chose to modify their IRSs in T1, whilst the majority did so to modify length, a statistically significant number also chose to modify its balance; with more categories being added to the 'agreement' side of the continuum. Paired sample t-tests showed that respondents who modified their IRSs in T1, kept the same IRS balance, as well as length, when completing the survey in T2. For those that did not modify their IRSs in T1, a small statistical difference was found in the IRSs they defined in T2, with the IRSs becoming more balanced. There were no observable relationships between respondents' gender and level of study and their choice of IRS

balance. However, a Kruskal-Wallis test indicated a difference existed between degree disciplines and IRS balance chosen. An examination of means showed that respondents from the Business discipline had a tendency to define IRSs with more categories on 'agreement' than those from the Physical Sciences and IT disciplines, who generally opted for balanced IRSs. Whilst conclusive associations between individual characteristics and IRS balance were not found on the IRSP group as a whole, when the sample was split by degree discipline a single association was found; with respondents from a Business discipline, there was a positive correlation between Affective Orientation and IRS balance (of medium strength). It was not possible to test for whether IRS length or balance choices were affected by mood, a temporary state, given there was no significant difference in mood between the two time periods.

There was a very small error rate for inappropriate IRS verbal labels (1% - 2%). There were a total of 26 different verbal labels chosen for 'agree', and 30 for 'disagree' 'Completely', 'Totally', 'Definitely', 'Absolutely' and 'Strongly' were repeatedly in the top five most popular verbal labels chosen. The majority of respondents opted for imbalanced verbal labels in T1 (58%). Of those that completed both surveys (T1 and T2) approximately 50% of them chose imbalanced verbal labels in T1, but this number fell to around 40% in T2. Some may have opted to choose identical verbal labels the second time. Respondents who opted to modify their IRSs generally did so to alter the length of their scale, less than 2% changed their verbal labels.

Both test groups that experienced using the IRSP *and* the LTRS (TG2 and TG3) rated the IRSP more favourably across the three feedback items which compared the two methods: The respondents indicated that they paid more attention to survey items when

using the IRSP over the LTRS; they believed their answers more correctly reflected their opinions when using the IRSP; and they preferred using the IRSP over the LTRS. Respondents from both test groups found it easy to design their own IRS (although a much higher mean rating for 'Ease' was reported with TG3).

## **Chapter 6. Discussion and Conclusions**

## 6. Discussion and Conclusions

### 6.1 Introduction

The overall research objective of this project was:

*To develop and test a measurement instrument capable of having respondents individualise their own rating-scales for use in online surveys.*

This chapter completes the discussion of the quantitative findings, combines this discussion with the key qualitative insights from the development phase, and relates them back to the research objective. The implications of the research study are then raised. Finally, the study limitations are presented and how they might be overcome in future research, along with a more general discussion of future research that could extend the findings from this study.

### 6.2 Measurement Model Fit: IRSP vs LTRS

Confirmatory factor analysis (CFA) was conducted to establish whether loose cross-validation was present between the IRSP and LTRS groups, across three of the measurement models: Affective Orientation (AO), Personal Need for Structure (PNS), Cognitive Style Indicator (CoSI). Given that, with the LTRS group, measurement of these psychometric scales was conducted using the very Likert-type rating-scales recommended by the scale authors, the baseline expectation would be that the data from the LTRS group would better fit all three of the measurement models. However, this was not the case.

With the AO measurement model, the fit indices indicated a less than adequate fit for both groups (IRSP and LTRS). One might expect this to be the case for the IRSP group alone, given a new method of measurement was used to capture the data on the AO

items. However, the model fit was similarly unsatisfactory for the LTRS group. The fit statistics were extremely similar for both groups, with both falling slightly short of the ideal cut-off values. Convergent validity of the model was, however, present for both groups. When the model diagnostics were examined to improve model fit it was clear that, on the whole, the data from both groups yielded similar results. For example, the weaker performing items had low factor loadings across both groups. The re-specified AO model yielded acceptable fit statistics for both groups, with only a marginally better fit for the LTRS. Even after re-specification, the IRSP method captured the measurement model in a similarly satisfactory fashion to that of the LTRS.

Loose cross-validation was also present across groups with the original PNS model. The fit statistics indicated a similar fit for both groups, and was very close to being acceptable. Whilst the re-specified PNS model yielded an acceptable fit for both groups, it was a marginally stronger fit for the IRSP group.

Although both groups demonstrated inadequate fit for the CoSI measurement model, the IRSP was marginally better than that for the LTRS. Moreover, factor loadings were generally higher for the IRSP group than they were for the LTRS. As with the other models, weaker loading items were similar in both groups. Convergent validity, although substandard for two out of the three factors, was marginally better with the IRSP. There was an adequate level of discriminant validity for both groups, although again, the IRSP fitted better than the LTRS here. The re-specified CoSI model yielded an acceptable fit for both groups, with similar fit statistics between the groups.

These results suggest that the IRSP, as an instrument, is equally capable of capturing data that replicates the psychometric measurement models obtained through pre-validated LTRS methods. This is a very important finding, as it indicates that the IRSP was successful in accurately capturing respondents' underlying traits. Where the LTRS data yielded a fit, so too did the IRSP data. Similarly, where the LTRS data yielded a problem with either the fit or with particular items, the IRSP did also. More importantly, in some cases, the IRSP data fitted the measurement models *better* than the LTRS data.

### **6.2.1 Measurement Model Re-specification**

When re-specifying the three measurement models to improve fit, weaker loading items were removed (based on factor loadings, modification indices, and residuals). These items were cross-referenced back to those which had been noted in the qualitative phase as potential 'problem' items. Of the six items removed from the AO model, AO\_11 had been noted during the CVP-RD interviews (Stage 3 of the qualitative phase). The item wording was confusing to some respondents. It was reassuring, therefore, that it was picked up through the CFA analysis. AO\_14 was another item which had been removed from the model, and had also been noted previously during the CVP-RD interviews and during the pilot test with MBA students (Stage 4 of the qualitative phase). This item included the word 'subtle', the meaning of which may have eluded some respondents. No other AO items were noted as potential problem items during the qualitative phase.

Of the three items removed from the PNS model, PNS\_08 had been noted by a single respondent from the pilot test. This was due to the fact that they did not understand the meaning of the term 'unpredictable'. Generally speaking, most respondents would have understood the term, given the large majority of the sample spoke English as their



mother-tongue (but the respondent in question did not). However, it may have been one of the weaker items because it relates to other unknown underlying constructs. This item is phrased, “I hate to be with people who are unpredictable”. Whilst it might load on a Personal Need for Structure scale, someone could, for example, still have an extremely high personal need for structure but find ‘hate’ to be a strong word. This person might rate a low score on this item. Item PNS\_11 was another item which had been removed from the measurement model. It had been noted by only one respondent in the pilot test, but again, this was due to a language problem given he had not understood the meaning of the word ‘uncomfortable’. The remaining item which was removed, PNS\_01, had not previously been noted by respondents in any of the previous stages.

Of the four items removed from the CoSI model, none had been noted as potential problem items during the qualitative phase and pilot test. The two CoSI items that had been noted in previous stages were CoSI\_10 and CoSI\_17. CoSI\_10 had been highlighted only by those whose mother-tongue was not English, with four respondents from the pilot test not understanding the word ‘meticulously’. CoSI\_17 was noted for very different reasons, by both native and non-native speakers of English. In the CVP\_RD interviews, some respondents reported that this item (“I like to extend boundaries”) was vague in meaning. Whilst it was clearly interpreted as ambiguous by some respondents, it loaded quite highly on the CoSI\_Creating factor (.801) and was not problematic within the modifications indices and residuals outputs.

Thus, as was expected, there was some overlap between items that had been noted as potential ‘problem-items’ during the development phase of testing, and those that were

subsequently removed after confirmatory factor analysis was conducted. The CFA will, naturally, have uncovered problematic items (unidentified in the qualitative phase) due to their insufficient ability to contribute to the measurement of the underlying construct. As such, this result was not surprising. Given these items had loaded weakly with the LTRS group (i.e. using the very LTRS designed for use with the psychometric scale), this makes a contribution to the literature on these scales. Whilst not central to the purpose of this thesis, these findings contribute to the knowledge base concerning affective orientation, personal need for structure and cognitive style indicators. The authors and users of these psychometric scales may find these results useful in identifying potentially problematic items. However, these findings must be taken in context; the British student population.

### **6.3 Testing Multi-Group Measurement Model Equivalence**

Whilst the re-specified models for all three psychometric constructs indicated acceptable model fit for both the IRSP and LTRS data, measurement equivalence was tested through tighter cross-validation procedures. The IRSP and LTRS data passed the test of tight cross-validation on the AO measurement model, demonstrating factor structure, factor loading and error variances equivalence. Both groups also possessed strong cross-validation for the PNS and CoSI models, demonstrating factor structure, factor loading, and inter-factor covariance equivalence with each. The presence of full factor loading equivalence between the groups meant that differences between the values obtained from either measurement method could be compared. This rendered all subsequent comparisons between the two groups (IRSP and LTRS) *valid*. An additional test of scalar equivalence could not be conducted given the ML estimation method

could not be employed. This would have confirmed whether or not the *scores* between the two types of rating-scales (IRSP and LTRS) had the same meaning.

#### **6.4 Test-Retest Reliability**

Experimental mortality was similar for both those that completed the IRSP and the LTRS in T1. This result was surprising given the survey using the IRSP method will have been approximately 6 minutes longer (on average, respondents took 5.45 minutes to define their IRSs) than the LTRS. Given the IRSP survey takes longer to complete, and that it could be considered more burdensome for respondents, one might have expected more respondents (proportionately) from the T1 LTRS group to return for a re-test. This result suggests that despite the fact that the IRSP survey takes slightly longer, it did not appear to discourage respondents from returning for a re-test. This could mean that the extra task, of defining an IRS, was not considered *burdensome*, and is encouraging when considering its future potential.

The test-retest reliability of the IRSP method was compared to the LTRS, across the eleven factors. These results did not conclusively favour either method, as both performed similarly well. The IRSP group achieved higher test-retest reliability for the AO factor, two out of the three CoSI factors, and two out of the five BFI factors. Whilst one method was not found to clearly out-perform the other in terms of test-retest reliability, the fact that the IRSP performed equally well, is an encouraging result. Given respondents' IRSs were a subjective creation, and therefore, potentially change over time for some, one may have expected a lower degree of test-retest reliability for this group. The LTRS will have provided respondents with a fixed rating tool, and so in T2, respondents will have been using an LTRS that was familiar to them and

unchanged. For this reason, a result where the LTRS possessed better test-retest reliability may not have been surprising. However, the IRSs defined and used by respondents clearly must have been personally meaningful to those respondents to yield such levels of stability. Furthermore, test-retest reliability was not worse for those who defined different IRS lengths between the two time periods. This result is particularly interesting. It means that the addition or removal of intervals from their IRSs did not affect the test-retest reliability of those respondents' scores. It would seem that whilst there are transient factors that result in a change in the number of ideal response categories desired, this does not impact upon the ability of this individualised tool to capture their traits. This is an important finding.

The results of the multitrait-multimethod matrix analysis (Campbell and Fiske, 1959) indicated that internal consistency was present for all of the measurement models captured using both methods, except for the five Big Five Inventory (BFI) factors. The BFI factors produced low coefficient alphas in both the IRSP and LTRS groups. It was considered that because each of the BFI factors consisted of two-item measures, they may have been more prone to internal consistency problems. This is regardless of the method of measurement used. Looking at the other constructs, the AO factor produced a particularly high coefficient alpha in both groups; .930 with the IRSP data, and .905 with the LTRS data. Given that this factor score was generated from measures on eleven items, this provides further reason to believe that the BFI coefficients may have been as low as they were because they were two-item measures. The coefficient alphas were acceptable for the PNS and CoSI factors, and were similar between the two groups. Whilst in some cases the IRSP quadrant yielded higher coefficients than in the LTRS quadrant, generally speaking figures were comparable between the two groups.

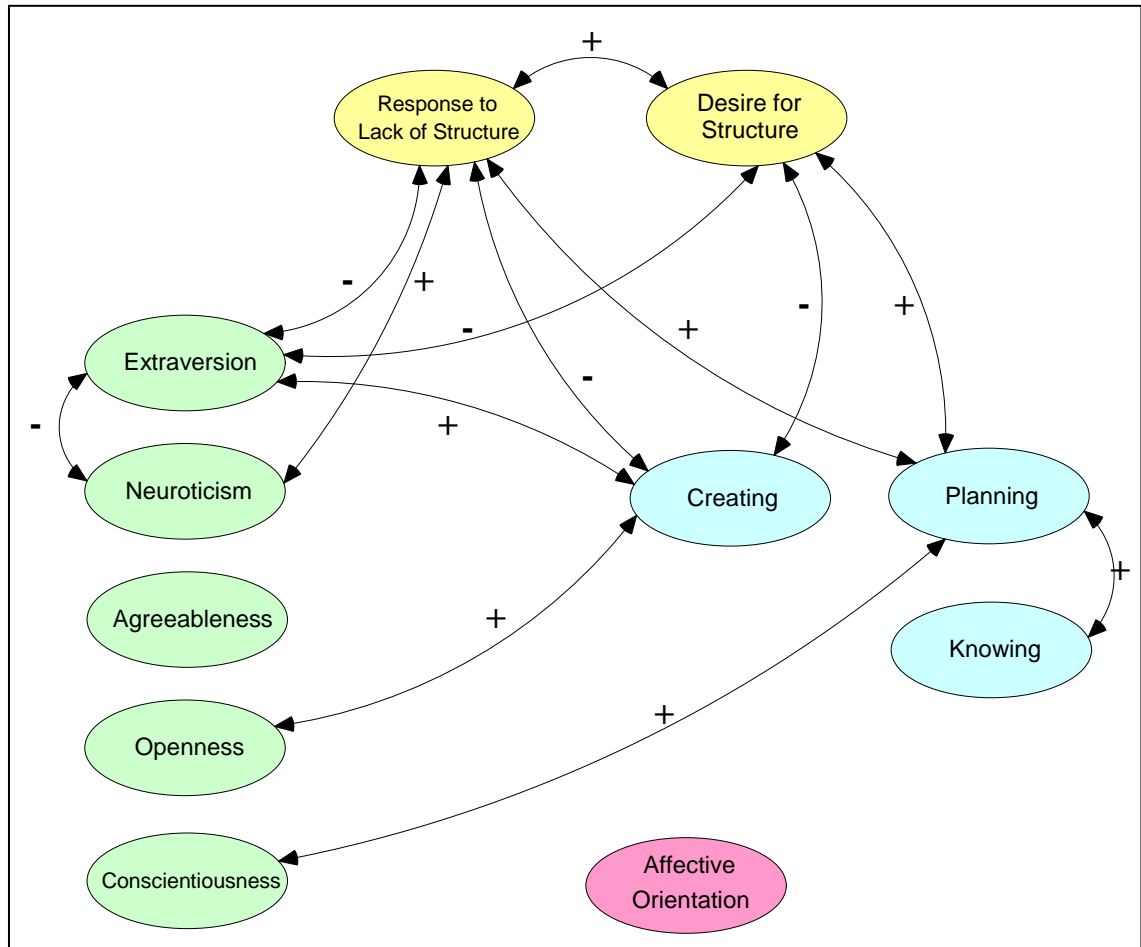
## 6.5 Further Test of Validity

The measures of the AO, CoSI and PNS traits achieved both convergent and discriminant validity under Campbell and Fiske's adapted method; that is, the utilisation of T2 data in the hetero-method quadrant. This final test of construct validity was necessary given Stage 1 had only established *within* model validity of each of the factors, and had not examined the validity of factors *between* models. This test extended the findings from Stage 1, and confirmed conclusively that convergent and discriminant validity was present, *between* groups (the mono-method quadrants; IRSP and LTRS) and *across* groups (the hetero-method quadrant). Whilst this test extended the test of validity, further confirming that the IRSP method can measure the psychometric constructs in an equivalent manner to the LTRS, it also provided information on the relationships between the constructs.

The patterns of correlations between the heteromethod quadrant and the two monomethod quadrants were very similar. In fact, at a cut-off of  $\pm 0.2$  or higher for the inter-factor correlations, the relationships are replicated across all the quadrants (Table 5. 42, Chapter 5). Where all but one correlation coefficient was significant but did not meet the cut-off value, this relationship was still included in the diagram where the figure was larger than  $\pm 0.150$  (Figure 6. 1).<sup>1</sup> This means that the IRSP managed to capture the same inter-factor relationships identified by the LTRS method.

---

<sup>1</sup>There were only four instances where this occurred. With the LTRS quadrant, the Knowing-Planning correlation is slightly under the cut-off at .186. With the IRSP quadrant: the Desire for Structure-Extraversion correlation is slightly under the cut-off at -.195; and the Creating-Extraversion correlation is slightly under the cut-off at .177. With the heteromethod quadrant, the Desire for Structure-Extraversion correlation is under the cut-off at -.165



**Figure 6. 1 Patterns of correlations identified through the MTMM matrix.**

The correlations identified are conceptually logical. Theoretically, this supports the relationships previously found in the literature, between both PNS and CoSI and the big five personality dimensions. For example, if one has a ‘desire for structure’ (DS), it is probable that one is less likely to be ‘creative’, given creativity is usually associated more with those who possess ‘flexible’ minds that see things in new ways. It is unsurprising therefore that ‘response to lack of structure’ (RLS) was also found to be negatively correlated with those adopting a ‘Creating’ cognitive style. Not only are the correlations logical, but they also support the findings of other studies:

- Neuberg and Newsom (1993) identified the following correlations with the PNS factors:

- RLS correlated positively with Neuroticism<sup>2</sup> ( $r = .32, p < .01$ ).
- RLS correlated negatively with Extraversion ( $r = -.23, p < .05$ ).
- Cools and Van den Broeck (2007) identified the following correlations with the CoSI factors:
  - Planning correlated positively with Conscientiousness<sup>3</sup> ( $r = .57, p < .01$ ).
  - Creating correlated positively with Extraversion ( $r = .24, p < .05$ ).

It is extremely encouraging that the results from the IRSP quadrant triangulate well with the results from the LTRS quadrant, and with the results of other studies. This shows that the IRSP is equally as capable of uncovering relationships between underlying traits as the LTRS.

Neuberg and Newsom (1993) also found a negative correlation between Desire for Structure (DS) and Openness ( $r = -.26, p < .01$ ). This relationship is not depicted in Figure 6. 1 because the LTRS quadrant (and the heteromethod quadrant) did not uncover it. In contrast, the IRSP quadrant did show a significant negative relationship between these two factors ( $r = -.216, p < .01$ ). This correlation met the condition of the  $-.2$  cut-off. In addition, these authors also found quite a strong negative correlation between RLS and Openness ( $r = -.44, p < .001$ ). Whilst none of the quadrants identified a significant relationship between RLS and Openness above the  $\pm.2$  cut-off, the IRSP quadrant narrowly missed it with a significant negative correlation of  $r = -.198, p < .01$ . The LTRS quadrant on the other hand yielded a smaller significant negative correlation of  $-.116$ , at the  $p < .05$  level. Given Neuberg and Newsom (1993) will have used a

---

<sup>2</sup> Neuberg and Newsom (1993) employed a 44-item version of the BFI scale (John et al., 1991).

<sup>3</sup> They measured the 'big five' personality factors using the Single-Item Measures of Personality (SIMP) scale, which consists of five bipolar items, with each representing one of the poles of the 'big five' factors (Woods and Hampson, 2005).

LTRS in their study, it is somewhat surprising that it was the IRSP and not the LTRS quadrant that best supported these additional relationships. It is worth highlighting that the authors used a larger measure of the BFI (John et al., 1991), which consisted of 44 items. This might suggest that the IRSP was slightly better than the LTRS at capturing underlying traits where the psychometric scale consists of fewer items.

## **6.6 Individualised Rating-Scales (IRSs) and Individual Characteristics**

### **6.6.1 IRSP length**

That the most popular IRS chosen consisted of seven intervals, with 30% of the participants opting for it, lends some credence to Miller's 'magic number seven' argument (Miller, 1956). Researchers have been employing seven-point Likert-type rating-scales (LTRSs) for some time, and it would seem fitting that this came out as the most popular IRS length. This would be an appropriate point to recall one of the insights from the qualitative phase. A respondent's familiarity with fixed rating-scales was considered for its effect on their IRS. It was theorised, for example, that those who are very familiar with seeing seven-point LTRSs may be predisposed to defining an IRS of  $-3 \leftarrow 0 \rightarrow 3$ . The drawback here could be that they might not be defining their true ideal rating-scale. However, given the IRSP instructions prompted respondents to be introspective when anchoring their IRS for personally meaningful scenarios, this should have been avoided. Additionally, because respondents were provided with an option to modify their IRS, the issue was likely to have been circumvented. As such, those who opted for seven-point IRSs are likely to genuinely possess that number of ideal response categories, rather than unthinkingly re-defining typical LTRSs they have already seen in other surveys.



That 54.3% of the sample opted for lengths longer than seven-points might indicate that researchers have not been maximising the information-transmitting capacity of respondents. This is an important finding, which questions the use of Miller's magic number seven as the benchmark. Where respondents have been capable of gradating their opinions on longer rating-scales, researchers could have benefited from the additional information provided. The data obtained would be of a higher quality. Not only has this been a missed opportunity, but 'forcing' respondents who possess longer ideal rating-scales to use a fixed rating-scale with fewer categories, might result in the incorrect cognitive mapping of their ideal categories onto the rating-scale and, as a result, biased scores (Hui and Triandis, 1989). Respondents would have had to cluster their ideal response categories onto the intervals provided by the rating-scale, which introduces scope for error in obtained responses. If such large numbers of respondents are capable of using rating-scales longer than seven-points, it would be beneficial to both respondents and the researchers if there were a facility for this to take place. In this way the IRSP provides respondents with the facility to map their ideal categories more accurately. As a consequence, researchers would benefit from an increase in data quality.

When examining relationships between personal characteristics and IRS length, few were found. There were no differences between males and females on the IRS lengths chosen. No differences were found between undergraduates and postgraduates either. There were also no differences between respondents from various degree disciplines and their choice of IRS length. This result was surprising given that, from the qualitative phase, there was a hint that there were potential differences in the way

people from different degree disciplines defined their IRS length. Recall the example of Interviewee 11 whose personally meaningful way of looking at agreement/disagreement was grounded in the discipline he was studying, namely environmental sciences. His degree discipline had exposed him to other types of rating-scales and it influenced the way he gradated his opinions about other concepts. However, this was clearly not the case for enough respondents; otherwise a significant relationship between degree discipline and IRS length would have been found. It was also considered that different degree disciplines would attract different types of students, with different ways of *thinking*. It was therefore thought likely that there *would* be differences in the way respondents from different degree disciplines defined IRS length. As such, this lack of differentiation between degree disciplines was unexpected.

Of the eleven psychometric factors measured under the four scales (AO, PNS, CoSI and BFI), only two factors had a statistically significant relationship with IRS length. These were CoSI's Knowing Style ( $r = .100, p < .05$ ) and Creating Style ( $r = .122, p < .01$ ). However, these relationships, particularly Knowing Style, were too weak to be of meaningful importance. Conceptually, a positive relationship between Creating Style and IRS length seems logical. Cools and Van den Broeck (2007) described those with a Creating Style as those who: are creative; like experimentation; see problems as opportunities and challenges; like uncertainty and freedom. Conceptually, it therefore seems reasonable that the more creative types who prefer uncertainty might choose to use novel rating-scale lengths and be free from the constraints of a more 'black and white' response. However, the resultant relationship was not strong enough to be conclusive.

### 6.6.2 IRS balance

Qualitative insights showed that respondents' conceptual regard for agreement and disagreement influenced their choices when defining a rating-scale. The interviews indicated that those who regard agreeing and disagreeing in a bipolar fashion (as opposite ends on the same continuum), are likely to choose a numerically balanced IRS (i.e. the same number of intervals either side of neutral). However, whilst most respondents had a bipolar view of agreeing/disagreeing, there were some that regarded them in a more unipolar fashion. These respondents saw both 'agreeing' and 'disagreeing' as different feelings (rather than as opposites), and they preferred to have differing numbers of meaningful intervals for each. The results from the quantitative study supported the findings of the qualitative phase. The majority, approximately 80%, of the participants opted for a perfectly balanced IRS (i.e. same number of intervals on both sides). Thus, even when given the choice, most respondents prefer a numerically balanced IRS for agreement/disagreement. Of those that preferred a numerically imbalanced IRS, the tendency was to have more intervals on the 'agreement' side of the continuum. On the whole, respondents were happy with the original IRS balance defined. Even for those who opted to modify their IRSs (after practicing using it on Greenleaf's items), few changed its *balance* whilst many changed their IRS's *length*. The few who did opt to modify its balance, generally added extra intervals to the 'agreement' side of the continuum. The reason for this is unknown. However, it could indicate that people are better able to gradate their positive feelings to a finer level than their negative ones.

There was no difference in the IRS balance scores from T1 to T2, for those who chose to modify their IRSs within T1. This within-survey learning appeared to have stuck with

them even through the wash-out period, which is in accord with the findings on IRS length. However, a slight difference in IRS balance was found from T1 to T2, for those who did not modify their IRSs in T1. This difference was only slight and was not strong enough to be conclusive. The small statistically significant difference suggested that respondents who did not modify their IRSs in T1, decided to have more balanced IRSs in T2 (even though their IRS lengths did not change). This may have meant that a small proportion of respondents who had defined imbalanced IRSs in T1, kept the same number of intervals in T2 but decided that they preferred a symmetrical rating-scale. A short measure of 'mood' had been included in the survey, to test whether a transient state like 'mood' had any impact on the desire for a balanced IRS. However, there was no difference in 'mood' between the two time periods, so it is unlikely that this was the cause of the IRS balance differences. Consequently, the reason for this slight difference in balance is unknown. It may be that an imbalanced rating-scale was a novelty for some in T1, but the novelty-effect wore off the second time they completed the survey and so they opted for a rating-scale that was less novel but more personally appropriate. Even if this were the case, it did not affect the measurement properties of the IRSP given the satisfactory reliability and validity findings.

When investigating possible relationships between personal characteristics and IRS balance, few were found. No differences were found between males and females on their IRS balance scores. No differences were found between undergraduates and postgraduates on their IRS balance scores. However, an inspection of group means suggested that there were differences between how respondents from different degree disciplines defined their IRS in terms of balance. Overall, those from IT and Physical Sciences defined perfectly balanced IRSs, whereas those from Business tended to define

IRSs with a greater number of categories for ‘agreement’. No relationships between IRS balance and any of the eleven psychometric traits were found. However, when the data was split by degree discipline, one relationship emerged; for respondents from the Business discipline, a positive correlation was found between Affective Orientation and IRS balance ( $r = .322, p < .05$ ). Business students with higher Affective Orientation scores, therefore, tended to have a greater number of response categories on the ‘agree’ side of their rating-scale.

Typically, researchers use fixed rating-scales and impose numeric choices, and thus the agreement/disagreement conceptualisation, on respondents. However, the IRSP allows respondents to define their own rating-scale, in a way that reflects *their* conceptualisation of agreement and disagreement (as bipolar symmetrical, or unipolar asymmetrical). Even though the IRSP had the neutral position anchored at 0, this suited both those with a unipolar and a bipolar conceptualisation. For example, should respondents feel that they have three meaningful levels of *disagreeing* and four meaningful levels for *agreeing*, they could define an imbalanced IRS of  $-3 \leftarrow 0 \rightarrow 4$ . In this scenario the neutral point still works well, as it represents the absence of either extreme (as is done with bipolar continuums) whilst still allowing a form of unipolar conceptualisation. This appears to be a clear advantage that the IRSP has over researcher-defined fixed rating-scales such as LTRSs.

### **6.6.3 Verbal labels**

It was clear from Stage 1 of the qualitative phase that some respondents had experienced difficulty with an early version of the IRSP instructions, which resulted in inappropriate verbal labels. For example, Interviewee 1 had asked whether the aim was

to “put a word that *meant* “agree”” (i.e. a synonym for ‘agree’) in the space provided. By the final stage of the qualitative phase this no longer appeared to be an issue due to improvements in the IRSP verbal anchoring instructions. The fact that there was only a 1-2% error rate<sup>4</sup> observed in the quantitative study with the verbal label anchoring, is a very positive result. Some skim-reading is inevitable with any survey. It is likely that the 1-2% error resulted from those respondents who skim-read the exercise. However, there might be scope for further improvements to the verbal labelling instructions, so that this error rate could be reduced even more.

At one point during the development phase, a facility which could provide respondents with a list of adverbs was considered. However, qualitative insights indicated that respondents experienced a sense of achievement with the rating-scale they defined, increasing their involvement in the survey process. Whilst the relative ease of a list of adverbs was posed, many argued that they did not feel the list was necessary. For example, Interviewee 12 argued that whilst some might have a more expansive knowledge of adverbs than others, everyone is capable of choosing verbal anchors that are personally meaningful to them when describing their maximum agreement/disagreement. She believed that a list of adverbs would have detracted from the personalised nature of the rating-scale. The fact that 98% of respondents, in the quantitative study, had no problem choosing adverbs supports this qualitative observation.

The objective of the IRSP was to help respondents to access what, for them, are meaningful verbal labels for both endpoints of their continuum. The verbal labels

---

<sup>4</sup> Approximately 1% for ‘agree’ and 2% for ‘disagree’

chosen are meant to represent a respondent's conceptual extreme for agreement/disagreement. There were two adverbs that were the most frequently chosen on both sides of the continuum; 'Completely' and 'Totally'. In addition, 'Absolutely', 'Definitely', 'Strongly', and 'Really', all featured in the five most popular. The fact that 'Strongly' was never in the three most popular is interesting, with 92% of respondents opting for other adverbs. This is a worrying finding for researchers who use the adverb 'Strongly' in verbal endpoints on fixed rating-scales, which seek to represent respondents' entire continuum. It would seem that, if given the choice, respondents are choosing a multiple of other words which are more personally meaningful to them. How they might respond to fixed verbal labels that are not personally appropriate is an interesting area of further debate.

Some of the qualitative observations from this study suggested that respondents' may have more positive or negative associations with certain adverbs and therefore feel as though they are better suited to one particular direction on the continuum. Qualitative observations suggested that 'Absolutely' might be an adverb that is associated more with 'agreement' than with 'disagreement'. However, respondents who chose this adverb for 'agree' and 'disagree' were 10% and 8.2% respectively, which would indicate little difference in its directional association. On the other hand, 'Definitely' was very popular as an anchor for 'agree' (11.5%), yet it was not within the five most popular labels when it came to 'disagree' in one of the time periods. In fact, only 3.2% of respondents chose it for 'disagree'. This indicates a possible link between certain adverbs and a respondent's tendency to assign them to a particular direction on the continuum.

An observation from the qualitative phase suggested that respondents with a bipolar view of agreement/disagreement may be more inclined to choose the same adverb for both sides of their rating-scale. Interestingly, the quantitative phase of the study showed that of those who completed the IRSP in both time periods, 53.5% of them chose differing verbal labels for their endpoints in T1. This might suggest that around half of respondents view ‘agreeing’ and ‘disagreeing’ as uniquely distinct emotions and not opposites on a continuum. However, in T2 only 39.6% of respondents had differing verbal labels. This indicates that some may have opted to change from imbalanced to balanced verbal labels the second time around. Perhaps the ‘novelty’ effect played a part. Whereby some respondents do in fact see ‘agreeing’ and ‘disagreeing’ as bipolar opposites but wanted to choose differing (yet personally meaningful) verbal labels because of the novelty of being allowed to do so. In T2, whilst many maintained their imbalanced verbal labelling, some decided to revert to balanced labels. This could mean that the unipolar/bipolar conceptual regard for agreement/disagreement could be independent of whether same/different labelling is chosen. At this juncture, it is worth reemphasising one of the observations from the qualitative phase, where all respondents who chose different verbal endpoints indicated that both of their endpoints (although different) represented, for them, their extremes on the agreement/disagreement cognitive continuum.

The abovementioned findings are particularly interesting, given that respondents have been shown to interpret standardised verbal anchors in different ways; with varying intensity and quality (Rohrmann, 2003). Rohrmann (2003) highlighted the disadvantages of using standardised verbal anchors for all respondents; asserting that they are of inferior measurement quality and are prone to cultural biases. His research



supports this very point, as he emphasised the need to create rating-scales using verbal anchors that reflect the cognitions of respondents. This is a key facility provided by the IRSP.

#### **6.6.4 The Need for a ‘Practice Routine’**

Approximately one fifth of respondents modified their IRS length after they had practiced using it on Greenleaf’s sixteen uncorrelated items and on being presented with the graph page prompt. This ‘practice routine’ appeared to be an important part of the exercise. It would seem that the facility for respondents to practice using their IRSs was crucial, given so many opted to modify length. However, few opted to modify their IRS balance. Moreover, a very small number of respondents, who opted to modify their IRSs, modified their verbal labels (1.7% for ‘agree’, 0.9% for ‘disagree’). It was clear, therefore, that the key advantage of having had this ‘practice routine’ was that respondents could better access their ideal number of ideal categories. The fact that respondents who experienced within-survey learning in T1 (i.e. changed their IRS length after the ‘practice routine’) repeated the use of their finalised T1 IRS lengths in T2, is an interesting finding. Given those who modified their IRSs in T1 kept those modified IRSs for T2, and did not choose to modify further in T2, it would seem that the ‘practice routine’ was less crucial in T2. This suggests that the ‘practice routine’ might only be required the very first time a respondent is asked to individualise a rating-scale. Once they are familiar with the process of individualising a rating-scale, and have done it before, they may no longer need a ‘practice routine’ in future surveys. This is important, given this would significantly reduce the additional time added to survey completion when the IRSP method is employed. Respondents could be asked at the beginning of a survey whether they have ever previously individualised a rating-scale,

and if they click ‘yes’, they could be ushered through a faster version of the process. If they click ‘no’ then they could be provided with a ‘practice routine’, followed by the option to modify.

### **6.6.5 IRSP Feedback**

The quantitative findings indicated that the IRSP was preferred over the LTRS in three areas where respondents were required to make a direct comparison (on items ‘Attention’, ‘Meaningful’, and ‘Preference’). A fourth area, ‘Ease’, required respondents to rate the level of ease they experienced when designing their IRS, and the results showed that the participants generally found it easy.

However, it was apparent that the mean ratings for Test Group 3 (TG3), LTRS-IRSP, seemed proportionately higher than those for Test Group (TG2), IRSP-LTRS. Further testing indicated that an order-effect may have impacted items ‘Ease’ and ‘Attention’. On the other hand, there might be other reasons for the apparent differences between these two groups. One reason for the IRSP receiving higher mean ratings from TG3 could be that respondents were using their IRSs to do the rating. In other words, when rating the four feedback items, TG3 respondents were doing so using their IRSs, and TG2 respondents will have rated them using the LTRS. This raises two points for consideration. The first is that respondents who used their IRS to rate the items (TG3) were better able to reflect their opinions (which both groups agree is the case, as evidenced by the positive mean ratings for ‘Meaningful’ in both test groups). This could mean that TG3’s mean ratings are more reflective of the strength of respondents’ preference for the IRS over the LTRS. However, the second point is that, in the very act of using the IRS to rate the items, respondents may have felt biased toward scoring it

favourably, rendering TG2's mean ratings more trustworthy. Even if only TG2's mean ratings are considered, the findings are still very encouraging despite the fact that the IRSP scored more modestly with this group; the mean ratings across all four items were still skewed in favour of the IRSP over the LTRS.

The results across these four feedback items are discussed next, followed by a section on the implications of the results.

#### *6.6.5.1 Meaningfulness*

The qualitative phase showed it is clearly advantageous for the quality of data capture that respondents are able to define IRSs that are personally meaningful, both in terms of verbal and numerical conceptualisation. Round 4 of the (Stage 1) interviews showed that the IRSP was developing as required, given that most of the interviewees appeared to be defining and using personally meaningful IRSs. Many of the interviewees, when asked how they came to choose their numerical anchors, demonstrated a thought-process that was purposeful and meaningful. The quantitative results clearly support the interpretation that respondents feel their IRSs are more personally meaningful than typical LTRSs. On inspection of TG2, the IRSP was particularly favoured over the LTRS when it came to the 'meaningfulness' of rating-scales to respondents, scoring a mean rating of .26 (on a scale of -1 to +1). TG3 produced a slightly higher result for the 'Meaningful' item, with a mean rating of .31. These results indicate that, overall, respondents felt that their answers more accurately reflected their opinions if they used the IRSP over an LTRS.

#### 6.6.5.2 *Attention*

Some of the findings from the qualitative phase indicated that the IRSP, as a measurement method, could potentially augment respondents' involvement in the survey process. Interviewee 12, for example, said that because she was able to design her own rating-scale, she felt more involved in the questionnaire process and paid more attention to her responses than she would do in typical surveys. The quantitative results appeared to support the interpretation that respondents pay greater attention to their ratings when using the IRS. On inspection of TG3, the IRSP scored particularly well against the LTRS when it came to the 'attention' respondents report to be giving survey items, scoring a mean rating of .24. This meant that, on average, respondents felt that they paid more attention to survey questions when using the IRS to rate their responses. However, TG2 produced a much lower result for this item, with a mean rating of .08. Whilst still significant at the .05 level, in favour of the IRSP, it is a much lower score relative to that for the TG3. It is worth reiterating that respondents in TG2 used the LTRS to rate their feedback on the IRSP. The disparity between the scores from the two groups might be due to the aforementioned order-effect or due to the influence of the method used during the rating. Nonetheless, the result appears to support the findings of the qualitative insights, albeit conservatively in the case of TG2.

#### 6.6.5.3 *Ease*

Insights from the qualitative phase suggested that whilst the earlier versions of the IRSP were quite complicated to understand, after several iterative stages of refinement, the general feedback was that the interviewees found the IRSP easy to follow and execute. Almost all the interviewees indicated that they only needed to read the instructions once, and had found them easy to understand. This is clearly supported by the

quantitative results where the IRSP scored very highly on ‘ease of use’. TG3 yielded a very large .63 mean score and TG2, in keeping with the more modest trend, still resulted in .16. This is a very positive result, given it is crucial that the IRSP be a measurement method that respondents can use easily, otherwise it would have very little real world practicality.

#### 6.6.5.4 Preference

‘Which of the two measurement methods was preferred by respondents?’ was a key research question, as it interrogates whether the time required to develop an IRS is worthwhile for the respondents. The quantitative results were in favour of the IRSP. This was comfortably the case in both test groups. TG2 produced a mean score of .17, and TG3 of .25. Therefore, it would seem that, on the whole, respondents *want* the additional task of individualising their own rating-scale. This supports some of the qualitative observations, whereby many appeared to *enjoy* the personalised nature of using their own IRS to rate survey items.

## 6.7 Methodological limitations of the study and future research

### 6.7.1 Nonnormality of the data distribution

Whilst the nonnormality of the data distribution obtained in the quantitative study is technically not a limitation, as it could not be controlled for, the limitations of the measures used to address it are raised. Several methods have been put forward as a means for handling nonnormality with Structural Equation Modelling (SEM), discussed in the quantitative analysis chapter, including methods such as bootstrapping, ML estimation (some argued it is robust to nonnormality), and adjusted  $\chi^2$ . However, it is worth highlighting that the advice on dealing with nonnormality in SEM is mixed, and

that “considerable research remains to be conducted to determine what the optimal estimation procedure is for a given set of conditions” (Schumacker and Lomax, 2004: 69). However, after considering the available recommendations, the ADF method of estimation was chosen as it was the most appropriate when all constraints and factors were considered (e.g. degree of nonnormality, sample size, and measurement model complexity). Had the degree of nonnormality not been quite as extreme, ML estimation might have been appropriate. This would have permitted the use of SEM (a longitudinal model) in Stage 3 of the analysis, where the reliability of the method was tested across time. The longitudinal model doubles the variables in each measurement model, and so despite the large sample obtained, there were still not enough sample units to run SEM with ADF estimation adequately for Stage 3. Furthermore, the use of ML estimation would have permitted a test of scalar invariance (whether the quantifiable meanings of the rating-scales were the same in both groups) in addition to the test of measurement invariance (which provided an indication as to whether or not people from different groups interpreted and used the rating-scales in the same way – i.e. whether differences between the values obtained could be compared across groups).

In this context, it should be noted that whilst authors of all three psychometric scales (Cognitive Style Indicator, Affective Orientation and Personal Need for Structure) applied confirmatory factor analysis to the data, none mentioned whether the data distribution met the normality assumption (Neuberg and Newsom, 1993, Booth-Butterfield and Booth-Butterfield, 1996, Cools and Van den Broeck, 2007). Furthermore, only Cools and Van den Broeck’s (2007) study mentioned the CFA estimation method used (Maximum Likelihood) when validating the CoSI scale. It is

therefore not known whether the distributions had similar skewness and kurtosis levels to those evidenced in this study.

When examining the literature for other studies that have employed these psychometric scales, the AO scale has only been used in a single study that had applied CFA to the data obtained, and there was no mention of whether the data distribution was normal or what estimation method was used (Sinclair et al., 2010). Furthermore, this study had used the longer 20-item version of AO and not the 15-item version. With regard to the CoSI scale, there were no published studies, aside from the authors', that have tested the scale. This is less surprising with the CoSI scale given it is still very new, however, this means that there is no additional evidence pertaining to its reliability, validity and whether the distribution of scores on the CoSI are typically normally distributed. The PNS scale on the other hand has been around a lot longer, and is therefore better established. There are a number of studies that have employed its use, but most do not highlight whether the PNS model fit the data, and whether the data was distributed normally (Moskowitz, 1993, Schaller et al., 1995, Weary et al., 2001, Landau et al., 2004, Hodson et al., 2010). One study, by Hess et al. (2005), did indicate that the skew distribution associated with the PNS measure was nonnormal and that log transformations were performed on scores before factor analysis was employed. The degree of this departure from normality was not specified. The sample group obtained in Hess et al.'s (2005) study consisted of 151 adults from the general population with a wide age spread (23-86). Whilst it was different to the sample on which the current study was based, it would seem that nonnormal data distribution on the PNS scale is not unusual. In summary, a limitation of this research is that greater consideration could have been taken during scale selection to locate well established psychometric scales

with more empirical evidence available on the distribution of scores. This could potentially have circumnavigated the issues brought on by the severe degree of nonnormality and the abovementioned limitations to analyses.

Whilst handling nonnormality with SEM measures continues to evolve, it might therefore be particularly useful if future research included the measurement of constructs which have already been found to generate normal distributions. This would increase the likelihood of the data collected being normally distributed, and would therefore permit the more powerful SEM analyses using ML estimation (assuming a sufficiently large sample were also obtained). Alternatively, future research could aim for an even larger sample size (beyond 2,000 units), such that even if methods such as ADF estimation are used, limitations on analyses resulting from nonnormality are less likely. It should be duly noted, however, the resultant risk of having overpowered statistical tests.

### **6.7.2 Re-specification of the measurement models**

Worthy of mention here is the fact that the scales AO, PNS and CoSI, all had to have items removed due to an inadequate model fit. The fact that the models did not demonstrate adequate fit in either group was a surprising result, given the scales had previously been validated with the exact LTRs used in this study. Even though all three constructs had previously been validated (i.e. were found to be valid and reliable), some of the items did not load well and needed to be removed. Consequently, the models had to be re-specified to achieve an acceptable fit. However, re-specifying a model does not involve hard and fast rules, but an inspection of a number of factors (e.g. modification indices, residuals, factor loadings) between each step. As such, it is



acknowledged that this process is also dependent to some degree on the skill of the researcher. Several factors may have contributed to the slightly inadequate model fit.

Firstly, this may have been a result of substantive differences between this sample and those used in other studies. Whilst all three scales had used a mixture of student samples and samples from the general population during development, the CoSI student sample consisted solely of Belgians. Furthermore, the CoSI student sample consisted of MBA students, who tend to be older. The PNS and AO scales were developed and validated on American samples, both students and adults from the general population. There may, therefore, be substantive nuances from the other samples that could not be replicated in this study, and may have impacted on model fit.

Secondly, context effects from scale ordering and item ordering may have contributed, given “context can alter how respondents map their judgments onto the response scale and how they edit their answers before reporting them” (Tourangeau, 1999: 119). Consideration of these possible contributing factors raises another possibility for future research. It might be useful for future studies to simulate, as much as possible, the substantive environment from which the psychometric scale was validated. Given the test is of the *method*, and not the psychometric scale, it would be useful to keep all substantive variables the same but for the measurement method used. Differences in measurement model fit, therefore, would be less likely to result from extraneous variables.

### **6.7.3 Positively worded IRSP feedback items**

The survey in time period 2 included four items that measured respondent feedback on the IRSP when compared to the LTRS. A limitation of the findings is that the four items were positively worded statements about the IRSP. Ideally, negatively worded statements should also have been included, which would have better accounted for respondents with a tendency to acquiesce, thus biasing the responses. However, they were not included as an additional increase in survey length would have potentially frustrated respondents and contaminated the data. As such, future research might include fewer items measuring psychometric constructs, and more items (both positively and negatively worded) that closely measure the respondents' experience of the IRSP.

## **6.8 Contribution and Implications**

This study focused on the development of a new research method that can be employed by survey researchers, this method is designed to increase survey data quality. As such, this increase in data quality is central to the contribution of the IRSP, and the direct theoretical and practical implications of the study are concerned with marketing research methodology. This study established the reliability and internal validity of the IRSP, which is essential before moving on to demonstrating external validity (McGrath and Brinberg, 1983). In addition, it has been established that respondents report that the IRSP: is easy to execute; increases their attention to the survey questions; provides a more accurate reflection of their opinion than the LTRS; and, is preferred over the LTRS. While this study has demonstrated that the measurement method works under controlled conditions (i.e. internal validity), demonstrating external validity would provide greater support for the usefulness of the IRSP. This would involve modifying

the substantive features of the study, such as having different respondent characteristics (including culture), different constructs, and could extend to different types of rating-scale (e.g. semantic differential). The implications of the research findings are considered in the context of their theoretical contributions to marketing research and, where appropriate, the practical implications they have for business, along with how future research could demonstrate external validity.

### **6.8.1 Contribution to marketing research methods**

#### *6.8.1.1 Cognitive Aspects of Survey Methodology*

Whilst no strong, significant relationships were found between the individual traits measured by this study (both demographic and psychographic) and the type of IRS defined, this research still contributes to the CASM (Cognitive Aspects of Survey Methodology) movement. A deeper understanding of how respondents answer survey questions was gained, and it helped inform the creation of the IRSP. A dynamic measurement method that incorporates individual cognitive differences between respondents is very much in accordance with the type of research that contributes to furthering this movement. In the literature review, the response process based on the cognitive paradigm was outlined, and included a step where a respondent potentially has to format their judgement to 'fit' the response alternatives provided. CASM recognises that an inappropriate response that does not accurately reflect the respondent's opinion results from this action. As such, a valuable contribution to CASM is that this step in the response process is bypassed. Respondents using the IRSP will have already created a rating-scale appropriate for them, and would therefore not need to 'format their judgement' to fit a fixed rating-scale. Seventy percent of respondents did not choose the 'standard' seven response categories (Miller, 1956). In this way, if the IRSP were used,

it would overcome one of the key problems in the current response process, ensuring that each respondent had a rating-scale that was neither too long to invite scale attenuation, or too short and risk extreme responding.

#### *6.8.1.2 The Imbalanced IRS*

The quantitative findings showed that a significant portion of respondents prefer to have an imbalanced rating-scale, whereby they have more intervals on one side of their continuum than on the other. Observations from the qualitative phase suggested that many people find that they can gradate their opinion more finely when it comes to one side of the pole over the other (e.g. agreement over disagreement). It is a major drawback of fixed rating-scales such as Likert, that this requirement is typically not accommodated. The fact that the IRSP provides the facility for imbalanced rating-scales is a big advantage, contributing to the improvement to data quality through the reduction of error caused by respondents being forced to use balanced rating-scales, when an imbalanced rating-scale may be more appropriate for them.

#### *6.8.1.3 Construct meaningfulness and response category meaningfulness*

Both response category meaningfulness (Viswanathan et al., 2004) and construct meaningfulness (Gibbons et al., 1999) have been shown to be related to the manifestation of response styles. Triandis and Marin (1983) suggested that respondents use more extreme scores when the issues addressed are more meaningful to them (i.e. the topic is more personally relevant). Therefore, many cultural groups demonstrate an increased tendency to use the midpoints of the rating-scale and a decreased tendency to use the extremes, particularly when the administered questionnaires were developed in other cultures. In a study using three scales that measured attitudes towards gender

roles, Gibbons, Hamby and Dennis (1997) systematically investigated the item meaningfulness hypothesis. For each item, the respondent first rated his or her opinion and then rated the meaningfulness of the item to him or her, personally. They found a significant correlation between the distance of the attitude rating from the midpoint on the rating-scale and the meaningfulness of the item. The results of a study conducted by Gibbons, Zellner et al. (1999) demonstrated that if items are personally less meaningful, respondents can be expected to take the middle road or to express less intense opinions or attitudes.

In the context of meaningfulness of response categories, Viswanathan et al. defined this as “the number of categories that individuals typically use in thinking about an attribute in such situations as making a choice or judgement,” (2004: 199). It could be argued that the meaningfulness of the item is linked with meaningfulness of the response categories. Should the item in question be particularly meaningful to the respondent (e.g. perhaps the construct of interest is something they know a great deal about), then it would seem quite reasonable to assume that this respondent may be able to think about the item in detail and thus, in finer gradations (i.e. a rating-scale with greater response categories). In a situation where the item is less meaningful to the respondent, the reverse might occur (i.e. they need less response categories to report their cognitive judgement). Following this line of reasoning, it could be argued that problems are likely to occur if the researcher standardises the rating-scale length across a research instrument, where different constructs may have different levels of meaningfulness to respondents. It could be theorised that fixed rating-scale lengths do not allow for respondents to accurately reflect their cognitions on more meaningful items, given that respondents have greater capacity for discrimination with meaningful constructs (Couch

and Keniston, 1960, Viswanathan et al., 2004). This in turn affects the nature of what is represented by each response category, given that the item's meaningfulness has been linked to the response categories' meaningfulness. From this, a worthwhile area for further exploration would be to see whether the 'ideal' number of response categories will increase with construct meaningfulness.

#### 6.8.1.4 *Stability of IRSs*

Given the experimental nature of this research, little is known about how stable an individual's IRS is likely to be. The quantitative results indicated that even when respondents changed the length of their rating-scales between time periods 1 and 2, the ratings obtained for the latent constructs were the same. It would be useful to understand the degree to which a person's IRS is fixed. Perhaps there is a *trait* versus *state* trade-off that impacts on the stability of a respondent's IRS. Already mentioned was the consideration that the respondent's reaction to the construct under study is linked to the degree to which they can gradate their opinion about that construct (i.e. the number of intervals defined). However, what if the construct being measured does not change? There might still be other factors that impact on the respondent's decision to define a different IRS. Whether these are related to inherent traits (that are yet to be discovered) and/or transient states is a question that needs exploring. Furthermore, the most important question is whether the measurement stability of a person's IRS is sound, even if attributes change (e.g. number of intervals or verbal endpoints). For example, if a respondent is surveyed about construct of interest X, in this instance they may have defined a  $-3 \leftarrow 0 \rightarrow 3$  rating-scale, yet one year later when surveyed about exactly the same construct, they might define a  $-4 \leftarrow 0 \rightarrow 4$  rating-scale. Even though their IRS has changed, the question is whether the stability of their ratings has remained

constant (assuming their opinion about the construct is unchanged). It might also, therefore, be useful to experiment with different wash-out periods in longitudinal studies. Answering these questions, and exploring further what impacts on respondents' choices on the attributes of their IRS, is an important area for future research.

#### 6.8.1.5 *External validity*

This study used a multigroup experimental design to test a measurement instrument capable of having respondents individualise their own rating-scales. The measurement properties of the Individualised Rating-Scale (IRS) were compared to the Likert-type rating-scale (LTRS). As such, a homogenous sample (across key demographics) was used to limit the impact of extraneous variables and to facilitate the demonstration of internal validity. However, the use of a homogenous group limits the generalisability of the method. For example, because a highly educated sample was used (i.e. the student population), this raises the question of whether a less educated population may need greater support in the use of the Individualised Rating-Scale Procedure (IRSP). To demonstrate the external validity of the method, future research could replicate the multigroup experimental design used here in other populations. For example, a study with a sample from the general British population could be carried out.

The generalisability of the method could also be extended by looking at how it performed with different cultural groups. As such, the study could be repeated in other cultures; where a sample from each culture be exposed to both methods in the manner done in this study. This would permit *between-culture* and *within-culture* comparisons. Additional research of this kind would address a multitude of unknowns, such as

whether the IRSP instructions can be understood and carried out by persons from different backgrounds, with different levels of education, and different ages.

## **6.8.2 Contribution to management research practice**

### *6.8.2.1 The preference for IRSP: Increased attention and meaningfulness of categories*

That respondents claim to pay more attention to survey questions when using the IRSP and they feel it more accurately reflects their opinions has several practical implications. It is likely that respondents' overall degree of involvement in a survey is increased, and in turn they are more likely to think carefully about their responses. This would improve the accuracy of responses beyond just reducing response bias, therefore enhancing data quality. This reduction in bias, and increase in accuracy, would further render the quantitative findings more trustworthy. Obviously, this would be of great value to business researchers. Should future research also prove the theoretical proposition that the IRSP minimises response style bias, this too would yield significant improvements to data quality and substantive conclusions.

### *6.8.2.2 Response Styles*

Part of the IRSP had respondents practice using their IRSs on Greenleaf's sixteen uncorrelated items. This proved a valuable part of the process for several reasons: respondents could ascertain the ease-of-use of their IRS; and respondents could reflect on their responses to the items, and consider whether their intervals were distinctly meaningful to them before proceeding. However, those who completed the survey using the LTRS were also given Greenleaf's sixteen uncorrelated items to rate as this helped to ensure that both surveys were of equivalent length. As such, although it was not within the scope of this study, the ratings obtained on those sixteen items from the



quantitative study could be used to conduct further research on response styles. Raised in the literature review was the argument that inappropriate cognitive mapping of ideal response categories onto fixed rating-scales is a major cause of response bias (e.g. Hui and Triandis, 1989). Whilst this additional area was not within the scope of this study, the data from Greenleaf's sixteen items could be used to compute scores for extreme responding, (dis)acquiescence, and mid-point responding. Respondents' response style scores could be compared across both measurement methods. Additionally, Item Response Theory (IRT) could be employed to examine the data at an item-level rather than at scale-level. IRT has been used to measure response style contamination in data in order to reduce measurement inequivalence (Candell and Hulin, 1987, Drasgow, 1987, Ellis, 1989, Hambleton and Swaminathan, 1985, Hulin et al., 1983, Hulin, 1987, Hulin et al., 1982). Methods for investigating response bias contamination, such as those examples given here, would be of interest in future research with the IRSP. Should the IRSP be found to reduce the manifestation of response bias, data quality would be significantly enhanced. Further research in this area would be extremely useful, with some, such as Baumgartner and Steenkamp (2001), maintaining that an important contribution of further research be the identification of response formats that suffer least from response style bias.

A different but related area of interest is the graph page, as a step within the IRSP. The qualitative phase uncovered that the graph page had a particularly interesting effect on respondents who appeared to adopt stylistic responding. After rating Greenleaf's items and being shown the graph page, stylistic responders seemed to become aware of their tendency, with one stating that he tried to answer all subsequent questions in a more honest manner. The graph page offers a key step in the IRSP process, for respondents to

pause to reflect on to the meaningfulness of each response category before proceeding. Perhaps it also renders them aware of any response biases they suffer from. The effect of seeing the distribution of their responses to Greenleaf's items and the way one responds to all subsequent questions, should be explored further. It would be useful to examine what effect this is having on respondents. Should it make them aware in such a way that they try to minimise this stylistic tendency on all subsequent ratings, this mechanism (the graph page) in isolation could improve the quality of data obtained from electronic surveys. The IRSP Survey Software currently has the facility to remove or include the graph page as required. This means that the platform is already available for future research to be done.

#### *6.8.2.3 Measuring other concepts*

The IRSP was developed as a measure of agreement/disagreement with survey items. It was considered to be a valuable starting point, given it enabled a direct comparison with LTRSs (which measures level of agreement), and because the measurement of degree of agreement/disagreement appears so frequently in surveys. However, a valuable area for further research would be to extend the IRSP for the measurement of other concepts. The instruction wording would need to be adapted to other chosen concepts, and tested further both qualitatively and quantitatively. Further research of this kind could demonstrate that respondents are capable of individualising different types of rating-scales, based on a plethora of concepts. This raises two key issues for future research; the inclusion of the 'neutral' position, and the types of rating-scales featured in an IRSP.

As mentioned previously in the review of the literature, is the argument that the neutral position should not be included in rating-scales based on concepts that do not possess a

natural neutral. For example, another frequently measured concept is ‘the degree of satisfaction’ with the subject (e.g. object or service) of an item. Assuming the respondent has experienced the subject of the item, it has been argued that a neutral position should not be provided here given the respondent either was or was not satisfied (Presser and Schuman, 1980). Cases such as this, where the researcher does not wish to provide a neutral position, raises the question of how the IRSP adapts in this context. This would be a useful area to examine. Perhaps, a binary question could be posed as a precursor, such as “Were you satisfied or dissatisfied with this product?” Upon selecting one of two possible responses, they could then be asked to adjust the rating-scale length to represent the number of intervals they feel they have for this unipolar scale (e.g. if they were to have chosen ‘satisfied’ in the binary instance, followed by a choice of how many stages of satisfaction they feel they can experience). This would mean that they would be individualising a unipolar rating-scale.

Related to this is how respondents might individualise different types of rating-scale. The IRSP developed in this study was based on the Likert-type rating-scale. However, it should be possible for respondents to individualise rating-scales that are similar, for example, to the semantic differential rating-scale. Respondents could be given the two bipolar adjectives to consider, and individualise the number of intervals between them. In this way, the rating-scale would have already provided a conceptual and verbal anchoring, but the respondent could numerically anchor the rating-scale themselves. An experiment of this kind would provide scope for other very useful comparative studies, such as the data quality obtained from the IRSP compared to that obtained from the semantic differential rating-scale. Ideas such as these are valuable avenues for further

research, given the need for the IRSP to be flexible in order for its usefulness to be maximised.

### **6.8.3 Implications for Business**

Following on from the discussion on the contributions of this research to marketing research methods and management research practice, the implications for businesses (that engage in research) are summarised next.

The overarching implication of the IRSP for business centres on the increase in data quality. Businesses often make critical decisions (e.g. which new markets to enter, how to develop products/services) based on the business intelligence gleaned from consumer research. As was already evidenced in the literature review, poor data quality has led to invalid conclusions and the associated business cost (e.g. opportunity cost, decisions about products/markets). With inappropriate fixed rating-scales often resulting in response style bias, which subsequently lead to a data quality issue whereby Type I and Type II error have been shown to manifest, businesses have not benefitted from the confidence that the findings are indeed valid and free from contamination. The IRSP through circumnavigating the issues of inappropriate rating-scale length, the manifestation of response style bias, and its impact to data quality, can augment the confidence that businesses have in the validity of the conclusions drawn. This is of tangible value to businesses trying to find answers to the tough consumer questions, and the operational decisions that follow.

### 6.8.3.1 *The IRSP and pilot studies*

In situations where researchers would still prefer to use fixed rating-scales, either due to concerns over survey length or for other reasons (e.g., paper based survey administration required), the IRSP could still be employed in pilot studies. The business researcher conducting such pilot studies would be able to see what rating-scale length is likely to be the most suitable for that population if the rating-scale had to be fixed. They could also see which verbal anchors are most associated with the conceptual endpoints for that target population. This means that the IRSP can be used as a mechanism, through pilot studies, to justify why a particular rating-scale, when fixed in a study, was the most appropriate by number of intervals, symmetry and verbal labels. This would improve the validity of the measurement choices taken by the researcher, with the associated increase to the quality of the data obtained from an otherwise less appropriate fixed rating-scale selection.

### 6.8.3.2 *The IRSP and cross-cultural studies*

The IRSP could also be used in cross-cultural studies. Here there are known differences in rating-scale length preferences and other biasing factors relating to rating-scale length. If an electronic form of data collection is not being used, extending the IRSP method's application in a pilot survey context to the cross-cultural research environment could still result in fixed rating-scales that are more appropriate to each culture within a study. Given the IRSP would be providing respondents with fixed conceptual anchors, before having them numerically and verbally anchor the rating-scales, calibration will have occurred across cultures even if the rating-scale lengths preferred varied by culture. For example, if the mode rating-scale length chosen for pilot study in culture X was seven, yet it was five for culture Y, the researcher would fix these two rating-scales

respectively but be confident in the knowledge that both cultures had calibrated their preferred rating-scales to fixed conceptual anchoring.

Furthermore, whilst the evidence for providing an imbalanced rating-scale in the UK context is not strong (only around 20% of respondents designed an imbalanced rating-scale in the study), there is evidence that suggests rating-scales with more positive categories than negative would be appropriate in some national/cultural contexts (Riordan and Vandenberg, 1994). Even if it was not possible to use the IRSP in the cross-cultural main survey (e.g., if not all samples had access to the required technology), using the IRSP at the piloting stage of a cross-cultural survey would establish whether an imbalanced rating-scale was needed with any of the cultural groups being sampled.

This will result in rating-scales that are more appropriate to the culture of the sample being measured, with the associated improvements to data quality. In this way, the IRSP would be very useful to cross-cultural business researchers whether or not they decide to use the method for either the pilot study, main survey, or both.

#### *6.8.3.3 The IRSP and sample size*

This study, through its contribution to CASM via the creation of a measurement tool that more accurately captures respondents 'true' opinion, has major implications for business researchers. The IRSP's potential to increase the accuracy of responses, and associated reduction in data error, could allow sample size to be reduced. In addition, both the qualitative and quantitative findings indicate that a bi-product of the IRSP is increased engagement with the survey process. This is valuable to management

researchers as, amongst other things, it means that respondents are: more likely to complete the survey; more likely to answer sensitive questions, and; more likely to pay attention to each item individually rather than succumb to any global effects. The first two of these will lead to better response rates. A consequence of this may be a slight reduction in the cost (or time taken) as fewer members of the target population will need to be contacted to achieve the desired sample size. The greater attention paid to each item is likely to lead to better quality data. This may also mean that a smaller sample size is feasible to achieve the same level of confidence in the survey results. The larger the sample size (in probability samples), the less sensitive a data set is to error. However, if data error can be shown to be reduced, the need for a larger sample size is reduced accordingly. If smaller sample sizes could be shown to be reliable, given a reduction in data error, this would therefore yield cost savings for businesses by reducing the expense involved with sampling.

#### *6.8.3.4 Construct meaningfulness and rating-scale length*

The fact that construct meaningfulness has been linked with response category meaningfulness, could mean that business researchers may want to bear in mind the degree of involvement their target sample is likely to have in the construct of interest. For example, for a business wanting to research attitudes towards feminine hygiene products, the constructs of interest (e.g. image, price) may be more meaningful to female respondents (who are likely to know more about the products and be more engaged) than male respondents (who may have had limited exposure to the product, purchasing it on behalf of a partner or limited recall of advertising). In this scenario, female respondents might be more likely to be able to gradate their opinions more finely (i.e. have longer rating-scales), whereas male respondents may have a much shorter

spectrum of positions on the subject. Therefore, businesses employing fixed rating-scales should consider the degree of knowledge or exposure the target sample has had to the construct of interest, how meaningful it is to them, and how involved they are likely to be, when deciding upon rating-scale length. This will ensure that they are maximising the information transmitting capacity of the measurement instrument without allowing for the introduction of unnecessary response style bias. Future research in this area could examine whether rating-scale length correlates with involvement in the construct of interest. This could be done by measuring the respondents' degree of involvement in the subject before they define their IRS.

#### *6.8.3.5 Online surveys*

Respondents reported a higher level of involvement in the survey when using the IRSP. This has implications for businesses engaged in online survey research that suffer from low response rates. Given respondents find the IRSP more engaging, using this method of measurement could potentially improve response rates in online business research, and may also reduce problems with missing responses. However, whether this would be a long-term increase in response rates or a short-term boost to response rates (through a novelty effect), is unknown and is a worthwhile area of further study.

## **6.9 Concluding Remarks**

This research study successfully developed and tested a measurement instrument capable of having respondents individualise their own rating-scales for use in online surveys. Not only was this method found to be reliable and valid (internally) when compared to a popular fixed rating-scale, the Likert-type, but its dynamic nature was able to cater to individual-level differences ignored by fixed-rating scales. Whilst no



significant relationships were uncovered between the specific individual characteristics examined here and IRS choices, this finding has still contributed to a deeper understanding of respondents. It also reinforces the fact that setting up hard and fast rules about which fixed rating-scales should be used, ignores the adverse impact of unknown individual-level differences on data quality. The IRSP is able to help researchers circumvent this problem, while we continue to learn more about the cognitions of survey respondents.

In light of the shift to a more respondent-centric approach to measurement, such as the CASM movement, this new method of measurement is a very practical tool for researchers in a plethora of research contexts, such as:

- Pilot studies used to pre-test the measurement method;
- Cross-cultural studies where data comparability is often a problem;
- Online survey environments;
- Environments that typically suffer from low response rates;
- Situations where large sample sizes are difficult or costly to obtain.

The potential for improvements to data quality is substantial. Should future research find that the IRSP significantly reduces response style bias, this would have immense implications for business researchers and would help change the way they approach measurement.

## **Glossary**

---

## Glossary of Acronyms and Abbreviations

<b>Acronym/ Abbreviation</b>	<b>Meaning</b>
ADF	Asymptotic Distribution Free
AGFI	Adjusted Goodness of Fit Index
Agr	Agreeableness – A Big Five personality dimension
AN(C)OVA	Analysis of (Co)Variance
AO	Affective Orientation
AVE	Average percentage of Variance Extracted
BFI/ BFI-10	Big Five Inventory scale
c.r.	Critical Ratio
CASM	Cognitive Aspects of Survey Methodology
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
C-OAR-SE	Rossiter's (2002) C-OAR-SE procedure for scale development: Construct definition, Object classification, Attribute classification, Rater identification, Enumeration
Con	Conscientiousness – A Big Five personality dimension
CoSI	Cognitive Style Indicator
CoSI-C	Cognitive Style Indicator – Creating style
CoSI-K	Cognitive Style Indicator – Knowing style
CoSI-P	Cognitive Style Indicator – Planning style
CR	Construct Reliability
CVP	Concurrent Verbal Protocol
CVP-RD	Concurrent Verbal Protocols-Retrospective Debrief
Df	Degrees of freedom
DS	Desire for Structure – sub-dimension of the PNS
ERS	Extreme Response Style
Ext	Extraversion – A Big Five personality dimension
GFI	Goodness-of-Fit Index

GLS	Generalised Least Squares
IRSB_1	Balance of individualised rating-scale first chosen by respondent
IRSB_2	Balance of individualised rating-scale after modification (if applicable)
IRSB_Used	Balance of individualised rating-scale used by respondent in the survey
IRSL_1	Individualised rating-scale length first chosen by respondent
IRSL_2	Modified individualised rating-scale length (if applicable)
IRSL_Used	Individualised rating-scale length used in survey
IRSP	Individualised Rating-Scale Procedure
IRSPr1	First round of interviews in the iterative development of the IRSP (IRSPr2 would be the second round of interviews, and so on)
IRSPv1	Version 1 of the IRSP software prototype
IRSPv2	Version 2 of the IRSP software prototype
IRSS	Individual Rating-Scales
IRT	Item Response Theory
KAI	Kirton Adaption Innovation scale
KU	Kurtosis
LTRS	Likert-Type Rating-Scale
MI	Modification Index
ML	Maximum Likelihood estimation
MTMM	Multi-Trait Multi-Method
N	Experimental notation: Number of respondents
Neu	Neuroticism – A Big Five personality dimension
$O_i$	Experimental notation: Observation or measurement of experimental variables
Ope	Openness – A Big Five personality dimension
PNS	Personal Need for Structure
(R)	Experimental notation: Indication that respondents were randomly allocated to experimental groups
RD	Retrospective Debrief
RLS	Response to Lack of Structure – sub-dimensions of PNS
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation

SEM	Structural Equation Modelling
SK	Skew
SRMR	Standardized Root Mean square Residual
T1, T2	Indication of either the first or second data collection period
T1_IRS Length	Length of individualised rating-scale in time period 1
T2_IRS Length	Length of individualised rating-scale in time period 2
TF	Totally Free
TG <sub>i</sub>	Test Group i (1 to 4)
TLI	Tucker-Lewis Index
ULS	Unweighted Least Squares
WLS	Weighted Least Squares
X <sub>I</sub>	Experimental notation: Exposure of respondents to the IRSP
X <sub>L</sub>	Experimental notation: Exposure of respondents to Likert-type rating-scales

## **References**

---

## References

- ADLER, N., CAMPBELL & LAURENT, A. (1989) In search of appropriate methodology: From outside the PRC looking in. *Journal of International Business Studies*, 20, 61-74.
- ALBAUM, G. & PETERSON, R. A. (1984) Empirical research in international marketing 1976-1982. *Journal of International Business Studies*, 15, 161-173.
- ALWIN, D. F. (1992) Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological Methodology*, 22, 83-118.
- AMEMIYA, Y. & ANDERSON, T. W. (1990) Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics*, 18, 1453-1463.
- ANGELMAR, R. & PRAS, B. (1978) Verbal Rating Scales for Multinational Research. *European Research*, 62-67.
- ANON (2006) Inside Research.
- ARCE-FERRER, A. J. & KETTERER, J. J. (2003) The effect of scale tailoring for cross-cultural application on scale reliability and construct validity. *Educational & Psychological Measurement*, 63, 484-501.
- AULAKH, P. S. & KOTABE, M. (1993) An assessment of theoretical and methodological developments in international marketing: 1980-1990. *Journal of International Marketing*, 1, 5-28.
- BACHMAN, J. G. & O'MALLEY, P. M. (1984) Yea-Saying, Nay-Saying, and Going to Extremes: Black- White Differences in Response styles. *Public Opinion Quarterly*, 48, 491-509.
- BAGOZZI, R. P. (1984) A prospectus for theory construction in marketing. *Journal of Marketing*, 48, 11-29.
- BAGOZZI, R. P. (1994) Measurement in Marketing Research: Basic principles of questionnaire design. IN BAGOZZI, R. P. (Ed.) *Principles of Marketing Research*. Massachusetts, USA, Basil Blackwell Ltd.
- BARDO, J. W. & YEAGER, S. J. (1982) Note on reliability of fixed-response formats. *Perceptual and Motor Skills*, 54, 1163-1166.
- BARDO, J. W. & YEAGER, S. J. (1982b) Consistency of response styles across types of response formats. *Perceptual and Motor Skills*, 55, 307-310.
- BARDO, J. W., YEAGER, S. J. & BURDSAL, C. A. (1985) Examination of response formats without anchoring items. *Perceptual and Motor Skills Journal*, 61, 287-297.
- BARKSDALE, H. C. & MCTIER-ANDERSON, L. (1982) Comparative marketing: A review of the literature. *Journal of Macromarketing*, 2, 57-62.
- BARON, H. (1996) Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, 69, 49-56.
- BARTRAM, D. (1996) The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology*, 69, 25-39.
- BATTLE, C., IMBER, S., HOEHN-SARIO, A., NASH, E. & FRANK, J. (1966) Target complaints as criteria of improvement. *American Journal of Psychotherapy*, 20, 184-192.

- BAUMGARTNER, H. & STEENKAMP, J.-B. E. M. (2001) Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, 38, 143-156.
- BENDIG, A. W. (1953a) The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology*, 37, 38-41.
- BENDIG, A. W. (1953b) Effect of amount of verbal anchoring and number of rating scale categories upon transmitted information. *Journal of experimental Psychology*, 46, 87-90.
- BENDIG, A. W. (1954a) Reliability and the number of rating scale categories. *The Journal of Applied Psychology*, 38, 38-40.
- BENDIG, A. W. (1954b) Transmitted information and the length of rating scales. *Journal of experimental Psychology*, 47, 303-308.
- BENSON, P. H. (1971) How Many Scales and How Many Categories Shall We Use in Consumer Research?--A Comment. *Journal of Marketing*, 35, 59-61.
- BENTLER, P. M. (1990) Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107, 238-246.
- BERG, I. A. & COLLIER, J. S. (1953) Personality and group differences in extreme response sets. *Educational & Psychological Measurement*, 13, 164-169.
- BERRY, J. W. (1980) Introduction to methodology. IN TRIANDIS, H. C. & LONNER, W. (Eds.) *Handbook of Cross-Cultural Psychology: Methodology*. Boston, Allyn and Bacon.
- BHALLA, G. & LIN, L. Y. S. (1987) Cross-cultural marketing research: A discussion of equivalence issues and measurement strategies. *Psychology and Marketing*, 4, 275-285.
- BISHOP, G. F. (1987) Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly*, 51, 220-232.
- BLALOCK, H. M., JR (1979) The presidential address: measurement and conceptualization problems: the major obstacle to integrating theory and research. *American Sociological Review*, 44, 881-94.
- BLESS, H., BOHNER, G., HILD, T. & SCHWARZ, N. (1992) Asking difficult questions: Task complexity increases the impact of response alternatives. *European Journal of Social Psychology*, 22, 309-312.
- BLOOM, M., FISCHER, J. & ORME, J. (1999) *Evaluating practice: Guidelines for the accountable professional*, Boston, Allyn & Bacon.
- BOLLEN, K. A. (1989) *Structural equations with latent variables*, New York, John Wiley.
- BOLLEN, K. A. & BARB, K. H. (1981) Pearson's R and Coarsely Categorized Measures. *American Sociological Review*, 46, 232-239.
- BOLLEN, K. A. & STINE, R. A. (1993) Bootstrapping goodness-of-fit measures in structural equation modeling. IN BOLLEN, K. A. & LONG, J. S. (Eds.) *Testing Structural Equation Models*. Newbury Park, CA, Sage Publications.
- BOLTON, R. N. (1993) Pretesting Questionnaires: Content analyses of respondents' concurrent verbal protocols. *Marketing Science*, 12, 280-303.



- BOND, G., BLOCH, S. & YALOM, I. (1979) The evaluation of a "target problem" approach to outcome measurement. *Psychotherapy: Theory, Research, and Practice*, 11, 48-54.
- BOOTH-BUTTERFIELD, M. & BOOTH-BUTTERFIELD, S. (1996) Using your emotions: Improving the measurement of affective orientation. *Communication Research Reports*, 13, 157-163.
- BOYCE, A. C. (1915) Methods of measuring teachers' efficiency. *Nat. Soc. Stud. Educ.*, 14.
- BRESNAHAN, M. J., OHASHI, R., LIU, W. Y., NEBASHI, R. & LIAO, C.-C. (1999) A Comparison of Response Styles in Singapore and Taiwan. *Journal of Cross-Cultural Psychology*, 30, 342-358.
- BROUGHTON, R. & WASEL, N. (1990) A text-stimuli presentation manager for the IBM PC with ipsatization correction for response sets and reaction times. *Behavior Research Methods, Instruments and Computers*, 22, 421-423.
- BROWNE, M. W. (1984) Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.
- BROWNE, M. W. & SHAPIRO, A. (1988) Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, 41, 193-208.
- BYRNE, B. M. (1995) Structural Equation Modeling. Concepts, Issues, and Applications. IN HOYLE, R. H. (Ed.) *Structural Equation Modeling: Concepts, Issues, and Applications*. Thousand Oaks, California, Sage Publications.
- BYRNE, B. M. (2001) *Structural Equation Modeling with AMOS. Basic Concepts, Applications, and Programming*, Mahwah, New Jersey, Lawrence Erlbaum Associates.
- CAMPBELL, D. T. & FISKE, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- CAMPBELL, D. T. & STANLEY, J. C. (1966) *Experimental and quasi-experimental designs for research*, Chicago, Rand McNally & Company.
- CAMPBELL, K. S. (1999) Collecting Information: Qualitative Research Methods for Solving Workplace Problems *Technical Communication*, 46, 532-545.
- CAMPBELL, N. R. (1928) *An Account of the Principles of Measurement and Calculation*, London, Longmans, Green.
- CANDELL, G. L. & HULIN, C. L. (1987) Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology*, 17, 417-440.
- CHAMPNEY, H. & MARSHALL, H. (1939) Optimal refinement of the rating scale. *Journal of Applied Psychology*, 23, 323-331.
- CHEN, C., LEE, S.-Y. & STEVENSON, H. W. (1995) Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 6, 170-175.
- CHERRYHOLMES, C. H. (1992) Notes on pragmatism and scientific realism. *Educational Researcher*, 14, 13-37.

- CHEUNG, G. W. & RENSVDL, R. B. (2000) Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling. *Journal of Cross-Cultural Psychology*, 31, 187-212.
- CHILDERS, T. L. & RAO, A. R. (1992) The Influence of Familial and Peer-Based Reference Groups on Consumer Decisions *The Journal of Consumer Research*, 19, 198-211.
- CHUN, K.-T., CAMPBELL, J. B. & YOO, J. H. (1974) Extreme response styles in cross-cultural research. A reminder. *Journal of Cross-Cultural Psychology*, 5, 465-480.
- CHURCHILL JR., G. A. (1979) A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16, 64-73.
- CLOSS, S. J. (1996) On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology*, 69, 41-47.
- COMLEY, P. (2007) Online Market Research. IN VAN HAMERSVELD, M. & DE BONT, C. (Eds.) *Market Research Handbook*. 5th ed. Chichester, John Wiley & Sons, Ltd.
- CONKLIN, E. S. (1923) The scale of values method for studies in genetic psychology. *Univ. Ore. Publ.*, 2.
- CONRAD, F. G. (1999) Customizing survey procedures to reduce measurement errors. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley and Sons, Inc.
- COOLS, E. & VAN DEN BROECK, H. (2007) Development and Validation of the Cognitive Style Indicator. *Journal of Psychology*.
- CORNWELL, J. M. & DUNLOP, W. P. (1994) On the questionable soundness of factoring ipsative data: A response to Saville and Wilson (1991). *Journal of Occupational and Organizational Psychology*, 67, 89-100.
- COUCH, A. & KENISTON, K. (1960) Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of abnormal and social psychology*, 60, 151-174.
- COUCH, A. & KENISTON, K. (1961) Agreeing response set and social desirability. *Journal of abnormal and social psychology*, 62, 175-179.
- COX III, E. P. (1980) The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research*, 17, 407-422.
- CRANDALL, J. E. (1973) Sex difference in extreme response style. *Journal of Social Psychology*, 89, 281-293.
- CRANDALL, J. E. (1982) Social interest, Extreme response style, and implications for Adjustment. *Journal of Research in Personality*, 16, 82-89.
- CRESWELL, J. W. (1999) Mixed methods research: Introduction and application. . IN CIZEK, G. J. (Ed.) *Handbook of Educational Policy*. San Diego, Academic Press.
- CRESWELL, J. W. (2003) *Research design: Qualitative, quantitative, and mixed methods approaches*, California, Sage Publications Inc.
- CRONBACH, L. J. (1946) Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494.
- CRONBACH, L. J. (1950) Further evidence on response sets and test design. *Educational & Psychological Measurement*, 10, 3-31.

- CUNNINGHAM, W. H., CUNNINGHAM, I. C. M. & GREEN, R. T. (1977) The Ipsative Process to Reduce Response Set Bias. *Public Opinion Quarterly*, 41, 379-384.
- CURRAN, P. J., WEST, S. G. & FINCH, J. F. (1994) The robustness of test statistics and goodness-of-fit indices in confirmatory factor analysis. *Manuscript submitted for publication*.
- DAHLSTROM, R. & NYGAARD, A. (1995) An exploratory investigation of interpersonal trust in new and mature market economies. *Journal of Retailing*, 71, 339-361.
- DAVIS, H. L., DOUGLAS, S. P. & SILK, A. J. (1981) Measure unreliability: A hidden threat to cross-national marketing research? *Journal of Marketing*, 45, 98-109.
- DAWAR, N. & PARKER, P. (1994) Marketing Universals: Consumers' Use of Brand Name, Price, Physical Appearance, and Retailer Reputation as Signals of Product Quality *The Journal of Marketing*, 58, 81-95.
- DE JONG, M. G., STEENKAMP, J.-B. E. M., FOX, J.-P. & BAUMGARTNER, H. (2008) Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research*, 45.
- DIAMANTOPOULOS, A., REYNOLDS, N. L. & SIMINTIRAS, A. C. (2006) The impact of response styles on the stability of cross-national comparisons. *Journal of Business Research*, 59, 925-935.
- DILLMAN, D. A. & SMYTH, J. D. (2007) Design Effects in the Transition to Web-Based Surveys. *American Journal of Preventive Medicine*, 32, S90-S96.
- DONNELLY, C. & CARSWELL, A. (2002) Individualized outcome measures: A review of the literature. *Canadian Journal of Occupational Therapy*, 69, 84-94.
- DOUGLAS, S. P. & CRAIG, C. S. (1983) *International Marketing Research*, Englewood Cliffs, N. J.: Prentice Hall.
- DRASGOW, F. (1987) Study of measurement bias of two standardized psychological tests. *Journal of Applied psychology*, 72, 19-29.
- DUNCAN, O. D. (1984) *Notes on Social Measurement: Historical and Critical*, New York, Russell Sage.
- DURVASULA, S., ANDREWS, J. C., LYSONSKI, S. & NETEMEYER, R. G. (1993) Assessing the Cross-National Applicability of Consumer Behavior Models: A Model of Attitude Toward Advertising in General. *The Journal of Consumer Research*, 19, 626-636.
- EISER, J. R. & HOEPFNER, F. (1991) Accidents, disease, and the greenhouse effect: effects of response categories on estimates of risk. *Basic and applied social psychology*, 12, 195-210.
- ELLIS, B. B. (1989) Differential item functioning: Implications for test translations. *Journal of Applied Psychology*, 74, 912-921.
- ELLIS, B. B. & KIMMEL, H. D. (1992) Identification of Unique Cultural Response Patterns by Means of Item Response Theory. *Journal of Applied Psychology*, 77, 177-184.
- ENGLAND, G. W. & HARPAZ, I. (1983) Some methodological and analytic considerations in cross-national comparative research *Journal of International Business Studies*, 14, 49-59.

- FAZIO, R. H., POWELL, M. C. & WILLIAMS, C. J. (1989) The Role of Attitude Accessibility in the Attitude-to-Behavior Process. *Journal of Consumer Research*, 16, 280-288.
- FERGUSON, L. W. (1941) A study of the Likert technique of attitude scale construction. *Journal of Social Psychology*, 13, 51-57.
- FERRANDO, P. J. (2000) Testing the Equivalence Among Different Item Response Formats in Personality Measurement: A Structural Equation Modeling Approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 7, 271 - 286.
- FINN, R. H. (1972) Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational & Psychological Measurement*, 32, 255-265.
- FISCHER, R. (2004) Standardization to Account for Cross-Cultural Response Bias. A Classification of Score Adjustment Procedures and Review of Research in JCCP. *Journal of Cross-Cultural Psychology*, 35, 263-282.
- FRENCH-LAZOVIK, G. & GIBSON, C. L. (1984) Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement*, 8, 49-57.
- GARNER, W. R. (1960) Rating scales, discriminability, and information transmission. *The Psychological Review*, 67.
- GIBBONS, J. L., HAMBY, B. A. & DENNIS, W. D. (1997) Researching gender-role ideologies internationally and cross-culturally. *Psychology of Women Quarterly*, 21, 151-170.
- GIBBONS, J. L., ZELLNER, J. A. & RUDEK, D. J. (1999) Effects of Language and Meaningfulness on the Use of Extreme Response Style by Spanish-English Bilinguals. *Cross-Cultural Research*, 33, 369 - 381.
- GLASER, B. G. & STRAUSS, A. L. (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Chicago, Illinois, Aldine.
- GRAESSER, A. C., KENNEDY, T., WIEMER-HASTINGS, P. & OTTATI, V. (1999) The use of computational cognitive models to improve questions on surveys and questionnaires. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley & Sons, Inc.
- GRAVETTER, F. J. & WALLNAU, L. B. (2004) *Statistics for the Behavioral Sciences*, Belmont, CA, Wadsworth/Thomson Learning.
- GREEN, P. E. & RAO, V. R. (1970) Rating Scales and Information Recovery--How Many Scales and Response Categories to Use? *Journal of Marketing*, 34, 33-39.
- GREEN, P. E. & RAO, V. R. (1971) A Rejoinder to 'How Many Scales and How Many Categories Shall We Use in Consumer Research?--A Comment'. *Journal of Marketing*, 35, 62.
- GREENLEAF, E. A. (1992a) Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29, 176-188.
- GREENLEAF, E. A. (1992b) Measuring extreme response style. *Public Opinion Quarterly*, 56, 328-351.

- GRIMM, S. D. & CHURCH, A. T. (1999) A Cross-Cultural Study of Response Biases in Personality Measures. *Journal of Research in Personality*.
- GROVES, R. M. (1999) Survey error models and cognitive theory of response behavior. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley and Sons, Inc.
- GUILFORD, J. P. (1954) *Psychometric methods*, McGraw-Hill Book Company.
- GURWITZ, P. M. (1987) Ipsative rescaling: An answer to the response set problem in segmentation analysis. *Journal of Advertising Research*, 27, 37-42.
- GUY, R. F. & NORVELL, M. (1977) The neutral point on a Likert scale. *The Journal of Psychology*, 95, 199-205.
- HAIR, J. F., ANDERSON, R. E., TATHAM, R. R. & BLACK, W. C. (1998) *Multivariate Data Analysis*, New Jersey, Prentice Hall.
- HAIR, J. F., BLACK, W. C., BABIN, B. J., ANDERSON, R. E. & TATHAM, R. R. (2006) *Multivariate Data Analysis*, New Jersey, Pearson Prentice Hall.
- HAIR, J. F., BUSH, R. P. & ORTINAU, D. J. (2003) *Marketing Research: Within A Changing Information Environment*, New York, McGraw-Hill Inc.
- HAMBLETON, R. & SWAMINATHAN, H. (1985) *Item response theory: Principles and applications*, Boston, Kluwer-Nijhoff.
- HAMILTON, D. L. (1968) Personality attributes associated with extreme response style. *Psychological Bulletin*, 69, 192-203.
- HANCOCK, G. R. & MUELLER, R. O. (2006) *Structural Equation Modeling: A Second Course*, Greenwich CT, Information Age Publishing Inc.
- HARKNESS, J. A. (2003c) Questionnaire translation. IN HARKNESS, J. A., VAN DE VIJVER, F. J. R. & MOHLER, P. P. (Eds.) *Cross-Cultural Survey Methods*. Hoboken, NJ., John Wiley and Sons, Inc.
- HARKNESS, J. A., MOHLER, P. P. & VAN DE VIJVER, F. J. R. (2003a) Comparative research. IN HARKNESS, J. A., VAN DE VIJVER, F. J. R. & MOHLER, P. P. (Eds.) *Cross-Cultural Survey Methods*. Hoboken, NJ., John Wiley and Sons, Inc.
- HARKNESS, J. A., VAN DE VIJVER, F. J. R. & JOHNSON, T. P. (2003b) Questionnaire design in comparative research. IN HARKNESS, J. A., VAN DE VIJVER, F. J. R. & MOHLER, P. P. (Eds.) *Cross-Cultural Survey Methods*. Hoboken, NJ., John Wiley and Sons, Inc.
- HARTLEY, J., TRUEMAN, M. & RODGERS, A. (1984) The effects of verbal and numerical quantifiers on questionnaire responses. *Applied Ergonomics*, 15, 149-155.
- HEIDE, M. & GRONHAUG, K. (1992) The impact of response styles in surveys: A simulation study. *Journal of the Market Research Society*, 34, 215-230.
- HERRMANN, D. J. (1999) Potential contributions of the CASM movement beyond questionnaire design: Cognitive technology and survey methodology. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley & Sons, Inc.
- HESS, T., OSOWSKI, N. & LECLERC, C. (2005) Age and Experience Influences on the Complexity of Social Inferences. *Psychology and Aging*, 20.

- HODSON, G., RUSH J. & MACINNIS, C. (2010) Cavalier humor beliefs facilitate the expression of group dominance motives. *Journal of Personality and Social Psychology*, 99.
- HOFSTEDE, G. (1980) *Culture's Consequences: International differences in work-related values*, Beverly Hills, CA: Sage.
- HORN, J. L. & MCARDLE, J. J. (1992) A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18, 117-144.
- HORTON, R. L. (1974) The Edwards Personal Preference Schedule and consumer personality research. *Journal of Marketing Research*, 11, 335-337.
- HU, L.-T. & BENTLER, P. M. (1995) Evaluating Model Fit. IN HOYLE, R. H. (Ed.) *Structural Equation Modeling: Concepts, Issues, and Applications*. Thousand Oaks, California, Sage Publications.
- HUI, C. H. & TRIANDIS, H. C. (1989) Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20, 296-309.
- HULIN, C. (1987) A psychometric theory of evaluations of item and test translations: Fidelity across languages. *Journal of Cross-Cultural Psychology*, 18, 115-142.
- HULIN, C., DRASGOW, F. & KOMOCAR, J. (1982) Applications of item response theory to analysis of attitude scale translations. *Journal of Applied psychology*, 67, 818-825.
- HULIN, C., DRASGOW, F. & PARSONS, C. K. (1983) *Item response theory: Applications to psychological measurement*, Homewood, IL: Dow Jones Irwin.
- IWAWAKI, S. & ZAX, M. (1969) Personality dimensions and extreme response tendency. *Psychological Reports*, 25, 31-34.
- JACOB, H. (1971) Problems of scale equivalency in measuring attitudes in American subcultures. *Social Science Quarterly*, 61-75.
- JACOBY, J. & MATTEL, M. S. (1971) Three-Point Likert Scales Are Good Enough. *Journal of Marketing Research (JMR)*, 8, 495-501.
- JAHODA, M., DEUTSCH, M. & COOK, S. W. (1951) *Research methods in social relations*, New York, The Dryden Press.
- JAVELINE, D. (1999) Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly*, 63, 1-28.
- JENKINS, G. D. & TABER, T. D. (1977) A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, 62, 392-398.
- JOBE, J. B. & MINGAY, D. J. (1991) Cognition and survey measurement: History and overview. *Applied Cognitive Psychology*, 5, 175-192.
- JOHN, O. P., DONAHUE, E. & KENTLE, R. J. (1991) *The "Big Five" Inventory: Versions 4a and 54*, Berkley, CA, University of California, Institute of Personality Assessment and Research.
- JOHNSON, D. R. & CREECH, J. C. (1983) Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 48, 398-407.
- KALTON, G., COLLINS, M. & BROOK, L. (1978) Experiments in wording opinion questions. *Applied Statistics*, 27, 149-161.
- KALTON, G., ROBERTS, J. & HOLT, D. (1980) The effects of offering a middle response option with opinion questions. *The Statistician*, 29, 65-78.

- KILPATRICK, F. P. & CANTRIL, H. (1960) *Self-Anchoring Scaling: A Measure of Individuals' Unique Reality Worlds*, Washington DC, The Brookings Institution.
- KLINE, R. B. (2005) *Principles and practice of structural equation modeling*, New York, The Guilford Press.
- KOERBER, A. & MCMICHAEL, L. (2008) Qualitative Sampling Methods: A Primer for Technical Communicators. *Journal of Business and Technical Communication*, 22, 454-473.
- KOMORITA, S. S. (1963) Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology*, 61, 327-334.
- KOMORITA, S. S. & GRAHAM, W. K. (1965) Number of scale points and the reliability of scales. *Educational & Psychological Measurement*, 4, 987-995.
- KRISHNAMURTY, G. B., KASOVIA-SCHMITT, P. & OSTROFF, D. J. (1995) *Statistics: An Interactive Text for the Health and Life Sciences*, Boston, Jones and Bartlett.
- LANDAU, M., JOHNS, M., GREENBERG, J., PYSZCZYNSKI, T., MARTENS, A., GOLDENBERG, J. & SOLOMON, S. (2004) *Journal of Personality and Social Psychology*, 87.
- LEHTO, M. R., HOUSE, T. & PAPASTAVROU, J. D. (2000) Interpretation of fuzzy qualifiers by chemical workers. *International Journal of Cognitive Ergonomics*, 4, 73-86.
- LEWIN, K. (1951) *Field theory in social science: Selected theoretical papers*, New York, Harper.
- LEWIS, N. A. & TAYLOR, J. A. (1955) Anxiety and extreme response preferences. *Educational and Psychological Measurement*, 15, 111-116.
- LIGHT, C. S., ZAX, M. & GARDINER, D. H. (1965) Relationship of age, sex, and intelligence level to extreme response style. *Journal of Personality and Social Psychology*, 2, 907-909.
- LIKERT, R. (1932) A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1-55.
- LIPSEY, M. W. (1990) *Design Sensitivity: Statistical Power for Experimental Research*, CA, Sage Publications.
- LITTLE, T. D. (2000) On the Comparability of Constructs in Cross-Cultural Research: A Critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology*, 31, 213-219.
- LONG, W. (1946) Review: Philosophy of Business. *The Philosophical Review*, 55, 300-302.
- LORR, M. & WUNDERLICH, R. A. (1980) Mood states and acquiescence. *Psychological Reports*, 46, 191-195.
- MAY, H. (2006) A Multilevel Bayesian IRT Method for Scaling Socioeconomic Status in International Studies of Education. *Journal of Educational and Behavioral Statistics*, 31, 63-79.
- MCGRATH, J. E. & BRINBERG, D. (1983) External validity and the research process: A comment on the Calder/Lynch dialogue. *Journal of Consumer Research*, 10, 115-124.

- MEREDITH, W. (1993) Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.
- MERRENS, M. R. (1971) Personality correlates of extreme response style: A function of method of assessment. *Journal of Social Psychology*, 85, 313-314.
- MILLER, G. A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- MILLER, J. (2006) Online Marketing Research. IN GROVER, R. & VRIENS, M. (Eds.) *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*. Thousand Oaks, Sage Publications, Inc.
- MINTZ, J., LUBORSKY, L. & CHRISTOPH, P. (1979) Measuring the outcomes of psychotherapy: Findings of the Penn psychotherapy project. *Journal of Clinical and Consulting Psychology*, 47, 319-334.
- MOOIJART, A. & BENTLER, P. M. (1991) Robustness of normal theory test statistics in structural equation models. *Statistica Neerlandica*, 45, 159-171.
- MORGAN, D. (1998) Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative Health Research*, 8, 362-376.
- MORRISON, J., LIBOW, J., SMITH, F. & BECKER, R. (1978) Comparative effectiveness of directive vs. nondirective group therapist style on client problem resolution. *Journal of Clinical Psychology*, 34, 186-187.
- MOSKOWITZ, G. (1993) Individual differences in social categorization: The influence of personal need for structure on spontaneous trait inferences. *Journal of Personality and Social Psychology*, 65.
- MULLEN, M. (1995) Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, 26, 573-601.
- MURPHY, LIKERT & GARDNER (1938) *Public Opinion and the Individual: A psychological study of student attitudes on public questions, with a retest five years later*, New York, Russell and Russell.
- NAKAO, M. A. & AXELROD, S. (1983) Numbers are better than words. Verbal specifications of frequency have no place in medicine. *The American Journal of Medicine*, 74, 1061-1065.
- NEUBERG, S. L. & NEWSOM, J. T. (1993) Personal Need for Structure: Individual differences in the desire for simple structure. *Journal of Personality and Social Psychology*, 65, 113-131.
- NORMAN, R. P. (1969) Extreme response tendency as a function of emotional adjustment and stimulus ambiguity. *Journal of Consulting and Clinical Psychology*, 33, 406-410.
- NUGENT, W. R. (2004) A validity study of scores from self-anchored-type scales for measuring depression and self-esteem. *Research on Social work practice*, 14, 171-179.
- O'MUIRCHEARTAIGH, C. (1999) CASM: Successes, failures, and potential. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley and Sons, Inc.



- O'MUIRCHARTAIGH, C. A., KROSNICK, J. A. & HELIC, A. (1999) Middle alternatives, acquiescence, and the quality of questionnaire data. *Annu. Meet. Am. Assoc. Public Opinion Res.* Fort Lauderdale FL.
- OGASAWARA, H. (2003) Correlations among maximum likelihood and weighted/unweighted least squares estimators in factor analysis. *Behaviormetrika*, 30, 63-86.
- ORY, J. C. & POGGIO, J. P. (1981) Response variation effects on affective measures. *Educational and Psychological Measurement*, 38, 625-634.
- OSGOOD, C. E., SUCI, G. J. & TANNENBAUM, P. H. (1957) *The measurement of meaning*, Urbana, University of Illinois Press.
- OSKAMP, S. (1977) *Attitudes and Opinions*, Englewood Cliffs, N. J.: Prentice Hall.
- PAULHUS, D. L. (1991) Measurement and control of response bias. *Measures of personality and social psychological attitudes*. Academic Press, Inc.
- PAYNE, S. L. (1951) *The art of asking questions*, Princeton, Princeton University Press.
- PEABODY, D. (1962) Two components in bipolar scales: direction and extremeness. *Psychological Review*, 69, 65-73.
- PETERS, D. L. & MCCORMICK, E. J. (1966) Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology*, 50, 92-96.
- PIDGION, N. (1996) Grounded Theory: Theoretical Background. IN RICHARDSON, J. T. E. (Ed.) *Handbook of Qualitative Research Methods for Psychology and the Social Sciences*. Leicester, British Psychological Society Books.
- POYNTER, R. (2007) Main developments and trends. IN VAN HAMERSVELD, M. & DE BONT, C. (Eds.) *Market Research Handbook*. 5th ed. Chichester, John Wiley and Sons, Ltd.,
- PRESSER, S. & SCHUMAN, H. (1980) The Measurement of a Middle Position in Attitude Surveys. *Public Opinion Quarterly*, 44, 70-85.
- PRESTON, C. C. & COLMAN, A. M. (2000) Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104, 1-15.
- RAMMSTEDT, B. & JOHN, O. P. (2007) Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 203-212.
- RIORDAN, C. & VANDENBERG, R. (1994) A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, 20, 643-671.
- ROBSON, C. (2002) *Real World Research*, Oxford, Blackwell Publishers Inc.
- ROCKWELL, T. (2004) Rorty, Putnam and the pragmatist view of epistemology and metaphysics. IN MALACHOWSKI, A. (Ed.) *Pragmatism*. London, Sage Publications.
- ROHRMANN, B. (2003) Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Australia University of Melbourne.
- RORER, L. G. (1965) The great response-style myth. *Psychological Bulletin*, 63, 129-156.
- ROSSITER, J. R. (2002) The C-OAR-SE procedure for scale development in marketing. *International Journal of research in Marketing*, 19, 305-335.

- RUGG, D. & CANTRIL, H. (1944) *The wording of questions*, Princeton, Princeton University Press.
- SARIS, W. E. (1988) *Variations in response functions: A source of measurement error in attitude research*, Amsterdam, Sociometric Research Foundation.
- SATORRA, A. (2001) Goodness of fit testing of structural equation models with multiple group data and nonnormality. IN CUDECK, R., DU TOIT, S. H. C. & SORBOM, D. (Eds.) *Structural Equation Modeling: Present and Future*. Lincolnwood, IL, Scientific Software International.
- SAVALEI, V. (2008) Is the ML Chi-Square Ever Robust to Nonnormality? A Cautionary Note With Missing Data. *Structural Equation Modeling*, 15, 1-22.
- SAVALEI, V. & BENTLER, P. M. (2006) Structural Equation Modeling. IN GROVER, R. (Ed.) *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*. Thousand Oaks, California, Sage Publications.
- SCHAEFFER, N. C. (1991) Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly*, 55, 395-423.
- SCHAEFFER, N. C. & PRESSER, S. (2003) The science of asking questions. *Annual Review of Sociology*, 29, 65-89.
- SCHALLER, M., BOYD, C., YOHANNES, J. & O'BRIEN, M. (1995) The prejudiced personality revisited: Personal need for structure and formation of erroneous group stereotypes. *Journal of Personality and Social Psychology*, 68.
- SCHERMELLEH-ENGEL, K., MOOSBRUGGER, H. & MÜLLER, H. (2003) Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, 8, 23-74.
- SCHOBBER, M. F. (1999) Making sense of questions: An interactional approach. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley & Sons Inc.
- SCHUMACKER, R. E. & LOMAX, R. G. (2004) *A Beginner's Guide to Structural Equation Modeling*, Mahwah, New Jersey, Lawrence Erlbaum Associates.
- SCHUMAN, H. & PRESSER, S. (1977) Question wording as an independent variable in survey analysis. *Sociological Methods and Research*, 6, 151-70.
- SCHWARZ, N. (1990) Assessing frequency reports of mundane behaviors: Contributions of cognitive psychology to questionnaire construction. IN HENDRICK, C. & CLARK, M. S. (Eds.) *Research methods in personality and social psychology review of personality and social psychology*. Beverley Hills, Sage.
- SCHWARZ, N. (1999) Cognitive Research into Survey Measurement: Its Influence on Survey Methodology and Cognitive Theory. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley & Sons Inc.
- SCHWARZ, N., BLESS, H., BOHNER, G., HARLACHER, U. & KELLENBENZ, M. (1991b) Response scales as frames of reference: The impact of frequency range on diagnostic judgements. *Applied Cognitive Psychology*, 5, 37-49.
- SCHWARZ, N., HIPPLER, H.-J., DEUTSCHE, B. & STRACK, F. (1985) Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, 49, 388-395.

- SCHWARZ, N., KNAUPER, B., HIPPLER, H.-J., NOELLE-NEUMANN, E. & CLARK, L. (1991a) Rating Scales. Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55, 570-582.
- SHERIF, C. W., SHERIF, M. & NEBERGALL, R. E. (1965) *Attitude and attitude change*, Philadelphia, W. B. Saunders.
- SI, S. X. & CULLEN, J. B. (1998) Response categories and potential cultural bias: Effects of an explicit middle point in cross-cultural surveys. *International Journal of Organizational Analysis*, 6, 218-230.
- SILLS, S. J. & SONG, C. (2002) Innovations in Survey Research: An Application of Web-Based Surveys. *Social Science Computer Review*, 20, 22-30.
- SINCLAIR, M., ASHKANASY, N. & CHATTOPADHYAY, P. (2010) Affective antecedents of intuitive decision making. *Journal of Management & Organization*, 20.
- SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (1999) *Cognition and Survey Research*, Toronto, John Wiley and Sons, Inc.
- SIRKEN, M. G. & SCHECTER, S. (1999) Interdisciplinary Survey Methods Research. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley & Sons, Inc.
- SMITH, P. B. (2004a) Acquiescent Response Bias as an Aspect of Cultural Communication Style. *Journal of Cross-Cultural Psychology*, 35, 50-61.
- SMITH, R., OLAH, D., HANSEN, B. & CUMBO, D. (2003) The effect of questionnaire length on participant response rate: A case study in the U.S. cabinet industry. *Forest Products Journal*, 53, 33-37.
- SMITH, T. W. (2003b) Developing comparable questions in cross-national surveys. IN HARKNESS, J. A., VAN DE VIJVER, F. J. R. & MOHLER, P. P. (Eds.) *Cross-Cultural Survey Methods*. Hoboken, NJ., John Wiley and Sons, Inc.
- SMITH, T. W. (2004b) Methods for Assessing and Calibrating Response Scales Across Countries and Languages. *Sheth Foundation/Sudman Symposium on Cross-National Survey Research*. Champaign/Urbana.
- SNYDER, M. & ICKES, W. (1985) *Personality and social behavior*, New York, Random House.
- STEENKAMP, J.-B. E. M. & BAUMGARTNER, H. (1998) Assessing Measurement Invariance in Cross-National Consumer Research. *The Journal of Consumer Research*, 25, 78-90.
- STEMBER, H. & HYMAN, H. (1949) How interviewer effects operate through question form. *International Journal of Opinion and Attitude Research*, 3, 493-512.
- STEWART, D. W. (1981) The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, 18, 51-62.
- STRACK, F. & MARTIN, L. L. (1987) Thinking, judging and communicating: A process account of context effects in attitude surveys. IN HIPPLER, H. J., SCHWARZ, N. & SUDMAN, S. (Eds.) *Social Information Processing and Survey Methodology*. New York, Springer-Verlag.

- STRAUSS, A. & CORBIN, J. (1998) *Basics of Qualitative Research. Techniques and procedures for developing grounded theory*, California, Sage Publications Inc.
- SYMONDS, P. M. (1924) On the loss of reliability in ratings due to coarseness of the scale. *Journal of experimental Psychology*, 7, 456-461.
- TASHAKKORI, A. & TEDDLIE, C. (1998) *Mixed methodology: Combining qualitative and quantitative approaches*, California, Sage Publications Inc.
- TAYLOR, K. L. & DIONNE, J.-P. (2000) Accessing Problem-Solving Strategy Knowledge: The Complementary Use of Concurrent Verbal Protocols and Retrospective Debriefing. *Journal of Educational Psychology*, 92, 413-425.
- TENOPYR, M. L. (1988) Artifactual Reliability of Forced-Choice Scales. *Journal of Applied Psychology*, 73, 749-751.
- THYER, B. A., PAPSDORF, J. D., DAVIS, R. & VALLECORSIA, S. (1984) Autonomic correlates of the subjective anxiety scale. *Journal of Behavior Therapy and Experimental Psychiatry*, 15, 3-7.
- TOURANGEAU, R. (1984) Cognitive sciences and survey methods. IN JABINE, T., STRAF, M., TANUR, J. & TOURANGEAU, R. (Eds.) *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, DC, National Academy Press.
- TOURANGEAU, R. (1999) Context effects on answers to attitude questions. IN SIRKEN, M. G., HERRMANN, D. J., SCHECTER, S., SCHWARZ, N., TANUR, J. M. & TOURANGEAU, R. (Eds.) *Cognition and Survey Research*. Toronto, John Wiley and Sons, Inc.
- TOURANGEAU, R. & RASINSKI, K. A. (1988) Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299-314.
- TRIANDIS, H. C. & MARIN, G. (1983) Etic plus emic versus pseudoetic: A test of a basic assumption of contemporary cross-cultural psychology. *Journal of Cross-Cultural Psychology*, 14, 489-500.
- TUCKER, L. R. & LEWIS, C. (1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.
- TUCKETT, A. G. (2004) Qualitative research sampling: the very real complexities. *Nurse Researcher*, 12, 47-61.
- TVERSKY, A. & KAHNEMAN, D. (1974) Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1127-1131.
- VAN DE VIJVER, F. J. R. (2003) Bias and substantive analyses. IN HARKNESS, J. A., VAN DE VIJVER, F. J. R. & MOHLER, P. P. (Eds.) *Cross-Cultural Survey Methods*. Hoboken, NJ., John Wiley and Sons, Inc.
- VAN DE VIJVER, F. J. R. & LEUNG, K. (1997) *Methods and data analysis for cross-cultural research*, Thousand Oaks, Sage Publications Inc.
- VAN HEMERT, D. A., VAN DE VIJVER, F. J. R., POORTINGA, Y. H. & GEORGAS, J. (2002) Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. *Personality and Individual Differences*, 33, 1229-1249.
- VANDENBERG, R. J. & LANCE, C. E. (2000) A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, 3, 4-70.

- VIJIER, V. D. & POORTINGA, Y. H. (1982) Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology*, 13, 387-408.
- VISWANATHAN, M., SUDMAN, S. & JOHNSON, M. (2004) Maximum versus meaningful discrimination in scale response: Implications for validity of measurement of consumer perceptions about products. *Journal of Business Research*, 57, 108-124.
- WALLSTEN, T. S., BUDESCU, D. V. & ZWICK, R. (1993) Comparing the calibration and coherence of numerical and verbal probability judgements. *Management Science*, 39.
- WEARY, G., JACOBSON, J., EDWARDS, J. & TOBIN, S. (2001) Chronic and temporarily activated causal uncertainty beliefs and stereotype usage. *Journal of Personality and Social Psychology*, 81.
- WELKENHUYSEN-GYBELS, J., BILLIET, J. & CAMBRÉ, B. (2003) Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items. *Journal of Cross-Cultural Psychology*, 34, 702-722.
- WENG, L.-J. (2004) Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability. *Educational & Psychological Measurement*, 64, 956-972.
- WERNER, H. (1948) *Comparative psychology of mental development*, New York, Follett.
- WEST, S. G., FINCH, J. F. & CURRAN, P. J. (1995) Structural Equation Models with Nonnormal Variables: Problems and Remedies. IN HOYLE, R. H. (Ed.) *Structural Equation Modeling: Concepts, Issues, and Applications*. Thousand Oaks, California, Sage Publications.
- WILDT, A. R. & MAZIS, M. B. (1978) Determinants of Scale Response: Label Versus Position. *Journal of Marketing Research*, 15, 261-267.
- WOODS, S. A. & HAMPSON, S. E. (2005) Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality*.
- WYER, R. S. & CARLSTON, D. E. (1979) *Social cognition, inference and attribution*, Hillsdale, NJ, Erlbaum.
- WYER, R. S. J. (1969) The effects of general response style on measurement of own attitude and the interpretation of attitude-relevant messages. *British Journal of Social and Clinical Psychology*, 8, 104-115.
- ZAX, M., GARDINER, D. H. & LOWY, D. G. (1964) Extreme response tendency as a function of emotional adjustment. *Journal of Abnormal and Social Psychology*, 69, 654-657.
- ZUMBO, B. D. & ZIMMERMAN, D. W. (1993) Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, 34, 390-400.